Handbooks in Finance
2



HANDBOOK of ASSET and LIABILITY MANAGEMENT Volume 1

THEORY and METHODOLOGY

Editors: S.A. Zenios and W.T. Ziemba



North-Holland

HANDBOOK OF
ASSET AND LIABILITY MANAGEMENT
VOLUME 1:
THEORY AND METHODOLOGY

HANDBOOKS IN FINANCE

2

Series Editor

WILLIAM T. ZIEMBA

Advisory Editors

KENNETH J. ARROW
GEORGE C. CONSTANTINIDES
HARRY M. MARKOWITZ
ROBERT C. MERTON
STEWART C. MYERS
PAUL A. SAMUELSON
WILLIAM F. SHARPE



ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

North-Holland is an imprint of Elsevier



HANDBOOK OF ASSET AND LIABILITY MANAGEMENT

VOLUME 1: THEORY AND METHODOLOGY

Edited by

S.A. ZENIOS
University of Cyprus, Nicosia, Cyprus
and
The Wharton Financial Institutions Center, Philadelphia, PA

W.T. ZIEMBA
University of British Columbia, Vancouver, Canada



ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

North-Holland is an imprint of Elsevier



North-Holland is an imprint of Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK

First edition 2006

Copyright © 2006 Elsevier B.V. All rights reserved

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+44) (0) 1865 843830; fax (+44) (0) 1865 853333; e-mail: permissions@elsevier.com. Alternatively you can submit your request online by visiting the Elsevier web site at http://elsevier.com/locate/permissions, and selecting *Obtaining permission to use Elsevier material*

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Chapter 9, appearing on pages 385–428, was previously published in "Finding the Edge: Mathematical Analysis of Casino Games" (2000), pp. 163–214, O. Vancura, J.A. Cornelius and W.R. Eadington (Eds.), Reno, NV: Institute for the Study of Gambling & Commercial Gaming, University of Nevada.

ISBN-13: 978-0-444-50875-1
ISBN-10: 0-444-50875-9
ISSN: 1872-0978

For information on all North-Holland publications
visit our website at books.elsevier.com

Printed and bound in The Netherlands

06 07 08 09 10 10 9 8 7 6 5 4 3 2 1

Working together to grow
libraries in developing countries
www.elsevier.com | www.bookaid.org | www.sabre.org
ELSEVIER BOOK AID International Sabre Foundation

INTRODUCTION TO THE SERIES

Advisory Editors:

Kenneth J. Arrow, Stanford University, George C. Constantinides, University of Chicago, Harry M. Markowitz, University of California, San Diego, Robert C. Merton, Harvard University, Stewart C. Myers, Massachusetts Institute of Technology, Paul A. Samuelson, Massachusetts Institute of Technology, and William F. Sharpe, Stanford University.

The Handbooks in Finance are intended to be a definitive source for comprehensive and accessible information in the field of finance. Each individual volume in the series presents an accurate self-contained survey of a sub-field of finance, suitable for use by finance and economics professors and lecturers, professional researchers, graduate students and as a teaching supplement. The goal is to have a broad group of outstanding volumes in various areas of finance.

William T. Ziemba
University of British Columbia

Publisher's Note
For a complete overview of the Handbooks in Finance Series, please refer to the listing at the end of this volume.

This page intentionally left blank

CONTENTS OF THE HANDBOOK

Chapter 1
Enterprise-Wide Asset and Liability Management: Issues, Institutions, and Models
DAN ROSEN and STAVROS A. ZENIOS

Chapter 2
Term and Volatility Structures
ROGER J.-B. WETS and STEPHEN W. BIANCHI

Chapter 3
Protecting Investors against Changes in Interest Rates
OLIVIER DE LA GRANDVILLE

Chapter 4
Risk-Return Analysis
HARRY M. MARKOWITZ and ERIK VAN DIJK

Chapter 5
Dynamic Asset Allocation Strategies Using a Stochastic Dynamic Programming Approach
GERD INFANGER

Chapter 6
Stochastic Programming Models for Asset Liability Management
ROY KOUWENBERG and STAVROS A. ZENIOS

Chapter 7
Bond Portfolio Management via Stochastic Programming
M. BERTOCCHI, V. MORIGGIA and J. DUPAČOVÁ

Chapter 8
Perturbation Methods for Dynamic Portfolio Allocation Problems
GEORGE CHACKO and KARL NEUMAR

Chapter 9
The Kelly Criterion in Blackjack Sports Betting, and the Stock Market
EDWARD O. THORP

Chapter 10
Capital Growth: Theory and Practice
LEONARD C. MACLEAN and WILLIAM T. ZIEMBA

This page intentionally left blank

PREFACE

God may not be playing dice with nature, according to Einstein's famous quip, but chance and spontaneity are unavoidable in the financial markets. This simple but profound fact is driven eloquently in Bernstein's (1996) book. Organizations operating in the financial markets—be they financial institutions, manufacturing or distribution firms, or service providers—cannot achieve their main goals of creating economic value unless they come to terms with the risks associated with the uncertainties of the financial markets (Harker and Zenios, 2000). The alignment of a firm's operations and tactics with its uncertain environment is today perceived as a key strategy for all businesses. It draws the attention and demands heavy investment of resources by chief executives and boards of directors worldwide. The management of risky assets and uncertain liabilities in an integrated, coherent, framework not only remains the core problem of financial institutions today, but it has spilled over to other business enterprises as well.

The foundations for addressing today's challenges were laid more than fifty years ago in the Doctoral dissertation of Harry M. Markowitz (Markowitz, 1952, 1987, 1991) at the University of Chicago. This work laid the foundations for modern finance and was recognized by a Nobel prize in Economics in 1990. The early use of Markowitz's optimization theories was in developing normative models for understanding the financial markets, and as theoretical tools in financial economics. Since the 1980s however this line of research also evolved from a theoretical tool of positive analysis to a practical tool for normative analysis (Zenios, 1993). Optimization models are today at the core of decision support systems for financial engineers. The drive to integrate multiple interrelated risk factors of the global enterprise brought to the fore the power of asset and liability management models. At the same time developments of large-scale numerical optimization techniques, advances in optimization models for planning under uncertainty, and the availability of user-friendly modelling languages, put optimization tools in the hands of researchers and practitioners alike. Wallace and Ziemba (2005) provide details on many publicly available codes capable of solving very large practical scenario based stochastic programming ALM models.

Thus, the tools of asset and liability management have flourished. And the symbiosis between optimization tools and financial decision-making is becoming more fertile as we enter the 21st century marked by business globalization, rapid technological changes, financial innovations, and increased volatility in the financial markets.

Needless to say, the optimization models used in asset and liability management have been extended significantly—and in many cases deviated substantially— from the way shown by the pioneers of the fifties. The use of multi-period stochastic programming being perhaps the single most noteworthy generalization of the early works (Ziemba and

Vickson, 1975, 2006). Indeed, the proliferation of models for practical asset and liability modeling has been vast and witnessed by the sample of research articles collected in Ziemba and Mulvey (1998) and Wallace and Ziemba (2005), and the discussion of the practical use of these models in Ziemba (2003).

It is therefore fitting that the series ***Handbooks in Finance*** devotes a handbook to Asset and Liability Management. What may come as a surprise is that it took two volumes to collect what we perceive as the essential contributions of the last fifty years. Volume 1 contains chapters that lay the theoretical foundations and develop the methodologies essential for the development of asset and liability management models. Volume 2 considers several diverse business settings and a chapter devoted to each discusses problem-specific issues and develops realistic asset and liability management models. While all applications are drawn from financial institutions, readers interested in other business settings will find in both volumes sufficient material to gain deep insights into the asset and liability management modeling of other types of enterprises. The coverage is broad both in methodology and applications with chapters on term and volatility structures, interest rates, risk-return analysis, dynamic asset allocation strategies in discrete and continuous time, the use of stochastic programming models, bond portfolio management and the Kelly capital growth theory and practice in Volume 1. Volume 2 discusses applications of ALM models in banking, insurance, money management, individual investor financial planning, pension funds and social security.

We would like to thank all the authors for contributing chapters that address some aspect of asset and liability modeling that goes beyond the authors' own research contributions to the field. Having asked leading researchers to contribute each chapter we have been able to present the state-of-the-art in the field, while no efforts were spared in making the chapters accessible to a wider audience and not being restricted to the cognoscenti. And when a chapter may err on the side of focusing somewhat more narrowly on a specific research direction—dictated by the authors' preferences—extensive bibliographies at the end point readers to the vast fields beyond.

We hope that this collection of chapters and their references will be an invaluable resource for practitioners and the regulators of financial institutions, for researchers in the fields of finance, economics, mathematical finance and financial engineering, scholars in optimization, and mathematicians, scientists, and both doctoral and masters students.

Stavros A. Zenios
University of Cyprus and The Wharton Financial Institutions Center

William T. Ziemba
University of British Columbia

References

Bernstein, P.L., 1996. Against the Gods: The Remarkable Story of Risk. John Wiley and Sons, New York.
Harker, P.T., Zenios, S.A. (Eds.), 2000. Performance of Financial Institutions: Efficiency, Innovation, Regulations. Cambridge University Press, Cambridge, UK.

Markowitz, H.M., 1952. Portfolio selection. Journal of Finance 7, 77–91.
Markowitz, H.M., 1987. Mean-Variance Analysis in Portfolio Choice and Capital Markets. Basil Blackwell, Cambridge, MA.
Markowitz, H.M., 1991. Portfolio Selection: Efficient Diversification of Investments, second ed. Blackwell, Oxford, UK.
Wallace, S.W., Ziemba, W.T. (Eds.), 2005. Applications of Stochastic Programming. SIAM–MPS, Philadelphia, PA.
Zenios, S.A., 1993. Financial Optimization. Cambridge University Press, Cambridge, UK.
Ziemba, W.T., 2003. The Stochastic Programming Approach to Asset Liability and Wealth Management. AIMR, Charlottesville, VA.
Ziemba, W.T., Mulvey, J.M. (Eds.), 1998. Worldwide Asset and Liability Modeling. Cambridge University Press, Cambridge, UK.
Ziemba, W.T., Vickson, R.G. (Eds.), 1975. Stochastic Optimization Models in Finance. Academic Press, San Diego, CA; second ed. World Scientific, Singapore, 2006.

This page intentionally left blank

CONTENTS

Introduction to the Series v
Contents of the Handbook vii
Preface ix

Chapter 1
Enterprise-Wide Asset and Liability Management: Issues, Institutions, and Models
DAN ROSEN and STAVROS A. ZENIOS 1
Abstract 2
1. Introduction 3
1.1. What is enterprise risk management 4
1.2. Example: Enterprise-wide view of credit risks in a bank 5
2. A conceptual framework for enterprise risk management 6
2.1. The management of a single line of business 7
2.2. The management of a business portfolio 9
2.3. Integrating design, pricing, funding, and capitalization 9
2.4. Components of enterprise risk management 10
2.5. Why is enterprise risk management important 15
3. Asset and liability management in enterprise risk management 17
3.1. Components of asset and liability management 17
4. Models for asset and liability management 19
References 21

Chapter 2
Term and Volatility Structures
ROGER J.-B. WETS and STEPHEN W. BIANCHI 25
Abstract 26
Keywords 26
1. Term structure 27
1.1. An example 27
1.2. BootStrapping 30
1.3. Nelson–Siegel and Svensson's extension 34
1.4. Maximum smoothness 36
1.5. Forward-rates via geometric programming 37
1.6. EpiCurves 38
1.7. A comparison for U.S. Treasury curves 51

2. Volatility structure 57
2.1. Setting the stage 57
2.2. Some tree-based valuation models 60
2.3. The EpiVolatility model 61
2.4. Implementation 62
2.5. Summary 65
References 67

Chapter 3
Protecting Investors against Changes in Interest Rates
OLIVIER DE LA GRANDVILLE 69
1. Basic concepts for valuation and immunization of bond portfolios in continuous time 73
1.1. The instantaneous forward rate 73
1.2. The continuously compounded spot rate 75
1.3. Introducing the missing link: The continuously compounded total return 77
1.4. Relationships between the total return, the forward rate and the spot rate 80
1.5. Theorems on the behavior of the forward rate and the total return 81
1.6. The spot rate curve as a spline and its corresponding forward rate curve 84
2. Immunization: A first approach 90
2.1. The continuously compounded horizon rate of return 91
2.2. A geometrical representation of the horizon rate of return 91
2.3. Existence and characteristics of an immunizing horizon 93
2.4. The Macaulay concept of duration, its properties and uses 94
2.5. A second-order condition 99
2.6. The immunization problem 100
3. Protecting investors against any shift in the interest rate structure—A general immunization theorem 102
3.1. Notation 102
3.2. Present values at time 0 104
3.3. Future values at time 0 105
3.4. Present values at time ε 105
3.5. Future values at time ε 106
3.6. Further concepts for immunization: the moments of order k of a bond and a bond portfolio 106
3.7. A general immunization theorem 109
3.8. The nature of the cash flows of an immunizing portfolio 118
4. Applications 118
4.1. The spot structures and their shifts 119
4.2. Building immunizing portfolios 122
4.3. Immunization results 124
4.4. How large should we set the immunization parameter K? 126
4.5. Infinity of solutions 128
4.6. How sensitive are immunizing portfolios to changes in horizon H? 130

4.7. How sensitive are immunizing portfolios to a change in the basket of available bonds? 131
5. Conclusion and suggestions 133
6. Notes to references 137
References 137

Chapter 4
Risk-Return Analysis
HARRY M. MARKOWITZ and ERIK VAN DIJK 139
Abstract 140
Keywords 141
1. Introduction 142
2. The "general" mean-variance model 144
3. Applications of the general model 146
3.1. Asset liability modeling 146
3.2. Factor models 147
3.3. Other constraints 148
3.4. Tracking error 148
4. Examples of mean-variance efficient sets 149
4.1. Critical lines and corner portfolios 149
4.2. Efficient EV and $E\sigma$ combinations 150
4.3. All feasible $E\sigma$ combinations 152
4.4. Possible features 153
5. Solution to the "general" mean-variance problem 156
5.1. Preliminaries 156
5.2. The critical line algorithm 156
5.3. Getting started 158
5.4. The critical line algorithm with upper bounds 161
5.5. The critical line algorithm with factor and scenario models of covariance 162
5.6. Notes on computing 165
6. Separation theorems 166
6.1. The Tobin–Sharpe separation theorems 166
6.2. Two-funds separation 169
6.3. Separation theorems not true in general 169
6.4. The Elton, Gruber, Padberg algorithm 170
6.5. An alternate EGP-like algorithm 171
7. Alternate risk measures 173
7.1. Semideviation 173
7.2. Mean absolute deviation (MAD) 175
7.3. Probability of loss and value at risk (Gaussian R_p) 176
7.4. Probability of loss and Value at Risk (non-Gaussian R_p) 178
7.5. Conditional value at risk (CVaR) 180
8. Choice of criteria 180
8.1. Exact conditions 180

8.2. Mean-variance approximations to expected utility 181
8.3. Significance of MV approximations to EU 184
9. Risk-return analysis in practice 184
9.1. Choice of criteria 185
9.2. Tracking error or total variability 186
9.3. Estimates for asset classes 187
9.4. Estimation of expected returns for individual equities 187
9.5. Black–Litterman 188
9.6. Security analyst recommendations 189
9.7. Estimates of covariance 189
9.8. Parameter uncertainty 190
10. Epilogue 192
References 193

Chapter 5
Dynamic Asset Allocation Strategies Using a Stochastic Dynamic Programming Approach
GERD INFANGER 199
Abstract 200
Keywords 200
1. Introduction 201
2. Approaches for dynamic asset allocation 204
2.1. Multi-stage stochastic programming 204
2.2. Stochastic dynamic programming 206
3. Single-period portfolio choice 207
4. Utility functions 209
5. A general approach to modeling utility 211
6. Dynamic portfolio choice 214
6.1. Dynamic stochastic programming and Monte Carlo sampling 215
6.2. Serially dependent asset returns 216
6.3. A fast method for normally distributed asset returns 217
7. Numerical results 217
7.1. Data assumptions 217
7.2. An investment example 222
7.3. The performance of dynamic strategies 234
7.4. Dynamic strategies for hedging downside risk 238
7.5. Downside risk protection at every period 241
7.6. Computation times 246
8. Comparison to multi-stage stochastic programming 247
Acknowledgements 248
References 248

Chapter 6
Stochastic Programming Models for Asset Liability Management
ROY KOUWENBERG and STAVROS A. ZENIOS 253
Abstract 254
1. Introduction 255
2. Stochastic programming 256
2.1. Basic concepts in stochastic programming 256
2.2. Stochastic programming model for portfolio management 261
3. Scenario generation and tree construction 267
3.1. Scenarios for the liabilities 267
3.2. Scenarios for economic factors and asset returns 270
3.3. Methods for generating scenarios 272
3.4. Constructing event trees 277
3.5. Options, bonds and arbitrage 283
4. Comparison of stochastic programming with other methods 287
4.1. Mean-variance models and downside risk 287
4.2. Discrete-time multi-period models 288
4.3. Continuous-time models 290
4.4. Stochastic programming 291
5. Applications of stochastic programming to ALM 291
6. Solution methods and computations 296
7. Summary and open issues 297
References 299

Chapter 7
Bond Portfolio Management via Stochastic Programming
M. BERTOCCHI, V. MORIGGIA and J. DUPAČOVÁ 305
Abstract 306
1. Introduction 307
2. The bond portfolio management model 311
3. Input data 315
4. Scenario reduction and scenario tree construction 320
5. Numerical results 321
6. Stress testing via contamination: Add worst-case scenarios 325
7. Conclusions 334
Acknowledgements 335
References 335

Chapter 8
Perturbation Methods for Dynamic Portfolio Allocation Problems
GEORGE CHACKO and KARL NEUMAR 337
Abstract 338
1. Introduction 339

2. General problem formulation 340
2.1. Investment opportunity set 341
2.2. Utility function 344
3. Exact solution for unit elasticity of intertemporal substitution 345
3.1. General results 345
3.2. Example 1: Time-varying expected returns (finite horizon) 349
3.3. Example 2: Time-varying expected returns (infinite horizon) 352
4. Approximate solution for general elasticity of intertemporal substitution 353
4.1. Perturbation around unit elasticity of substitution 353
4.2. Perturbation around mean of consumption/wealth ratio 360
5. Example 366
5.1. Time-varying volatility 367
5.2. Time-varying interest rates 378
6. Conclusions 382
References 383

Chapter 9
The Kelly Criterion in Blackjack Sports Betting, and the Stock Market
EDWARD O. THORP 385
Abstract 386
Keywords 386
1. Introduction 387
2. Coin tossing 388
3. Optimal growth: Kelly criterion formulas for practitioners 392
3.1. The probability of reaching a fixed goal on or before n trials 392
3.2. The probability of ever being reduced to a fraction x of this initial bankroll 394
3.3. The probability of being at or above a specified value at the end of a specified number of trials 395
3.4. Continuous approximation of expected time to reach a goal 396
3.5. Comparing fixed fraction strategies: the probability that one strategy leads another after n trials 396
4. The long run: when will the Kelly strategy "dominate"? 398
5. Blackjack 399
6. Sports betting 401
7. Wall street: the biggest game 405
7.1. Continuous approximation 406
7.2. The (almost) real world 409
7.3. The case for "fractional Kelly" 411
7.4. A remarkable formula 414
8. A case study 415
8.1. The constraints 416
8.2. The analysis and results 416

8.3. The recommendation and the result 417
8.4. The theory for a portfolio of securities 418
9. My experience with the Kelly approach 419
10. Conclusion 420
Acknowledgements 420
References 428

Chapter 10
Capital Growth: Theory and Practice
LEONARD C. MACLEAN and WILLIAM T. ZIEMBA 429
Abstract 430
Keywords 431
1. Introduction 432
2. Capital accumulation 434
2.1. Asset prices 435
2.2. Decision criteria 435
2.3. Timing of decisions 436
3. Asset prices 436
3.1. Pricing model 438
3.2. Estimation 440
3.3. Comparison 442
4. Growth strategies 444
4.1. The Kelly strategy 445
4.2. Stochastic dominance 449
4.3. Bi-criteria problems: Fractional Kelly strategies 451
4.4. Growth-security trade-off 457
5. Timing of decisions 463
5.1. Control limits 463
6. Legends of capital growth 465
6.1. Princeton Newport Partners 466
6.2. Kings College Chest Fund 466
6.3. Berkshire–Hathaway 468
6.4. Hong Kong Betting Syndicate 469
References 469

Author Index 475
Subject Index 483

This page intentionally left blank

Chapter 1

ENTERPRISE-WIDE ASSET AND LIABILITY MANAGEMENT: ISSUES, INSTITUTIONS, AND MODELS*

DAN ROSEN

The Fields Institute for Research in Mathematical Sciences, 222 College Street, Toronto, Ontario M5T 3J1, Canada

STAVROS A. ZENIOS

HERMES Center on Computational Finance & Economics, University of Cyprus, P.O. Box 20537, 1678 Nicosia, Cyprus,
and
Financial Institutions Center, The Wharton School, Philadelphia, USA
e-mail: zenioss@ucy.ac.cy

Contents

Abstract 2
1. Introduction 3
   1.1. What is enterprise risk management 4
   1.2. Example: Enterprise-wide view of credit risks in a bank 5
2. A conceptual framework for enterprise risk management 6
   2.1. The management of a single line of business 7
      2.1.1. Designing the product 8
      2.1.2. Pricing the product 9
      2.1.3. Funding the product 9
   2.2. The management of a business portfolio 9
   2.3. Integrating design, pricing, funding, and capitalization 9
   2.4. Components of enterprise risk management 10
      2.4.1. Risk measurement 12
      2.4.2. Risk management 13
      2.4.3. Performance measurement 14
      2.4.4. Corporate governance 15
   2.5. Why is enterprise risk management important 15
3. Asset and liability management in enterprise risk management 17

* This work was completed while the second author was with Algorithmics Inc. as a Marie Curie Fellow of the European Commission. Research partially supported by EC grant ICA1-CT-2000-70015.

Handbook of Asset and Liability Management, Volume 1
Edited by S.A. Zenios and W.T. Ziemba
Copyright © 2006 Elsevier B.V. All rights reserved
DOI: 10.1016/S1872-0978(06)01001-5

3.1. Components of asset and liability management 17
4. Models for asset and liability management 19
References 21

Abstract

Enterprise risk management aligns a firm's business strategy with the risk factors of its environment in pursuit of business objectives. It is considered a well-grounded management strategy for corporations. The management of assets and liabilities is at the core of enterprise risk management for financial institutions. In this chapter we discuss the general framework for enterprise risk management, and the role of asset and liability management within this broader strategy. From the general concepts we proceed to focus on specific financial institutions, and conclude with a discussion of modelling issues that arise in the enterprise-wide management of assets and liabilities.

1. Introduction

> The revolutionary idea that defines the boundary between modern times and the past is the mastery of risk: the notion that the future is more than a whim of the gods and that men and women are not passive before nature.
>
> Peter Bernstein, *Against the Gods*.

Strategy is conceptualized as the alignment of a firm with its environment (Andrews, 1971 and Porter, 1980). *Enterprise risk management* (ERM) can then be defined as the strategy that aligns the firm's business with the risk factors of its environment in the pursuit of business goals.

Broadly speaking an organization's goal is to create economic value. Market value added is a generally accepted performance metric. However, this metric is difficult to measure and, furthermore, it can not be directly acted upon by management. Some performance metrics that are directly observable are growth and consistency of earnings, stability of cashflows, and the cost of financial distress. There is some evidence that these metrics are proxies for market value added. Although they are easily measured it is still the case that they can not be directly acted upon.

For the service industry in general, and financial institutions in particular, the identification of measures of performance that can be acted upon remains an open issue (Harker and Zenios, 2000). It is not obvious which managerial actions will improve consistency and growth of earnings, ensure cashflow stability, and reduce the cost of financial distress. A general framework linking strategy, processes, organizational forms, human resource management, information technology, and so on, to improvements in performance is not yet available. Such a framework would identify the *drivers of performance*, thus charting a path for specific managerial actions to improve the performance metrics. While a general framework remains elusive, there is some evidence that enterprise risk management is a driver of performance. A worldwide survey of insurance firms executives, conducted by Tillinghast-Towers Perrin, shows that a significant percentage (80%) consider ERM useful in improving earnings growth and consistency, and in achieving other metrics perceived by the executives as indicators of performance such as revenue growth, return on capital, expense control, etc. (Miccolis and Shah, 2000). It is fitting that ERM receives today attention from both academia and industry, although a stronger case still needs to be made that ERM is a significant driver of performance with measurable effects on the bottom line.

In this chapter we define ERM, describe a conceptual framework for an ERM strategy, and touch upon organizational issues. Asset and liability management (ALM) is then identified as a core activity of ERM for financial institutions. The chapter then proceeds to the discussion of asset and liability management for various financial institutions, and an overview of tools to support asset and liability management activities. The chapter concludes with an overview of the remaining chapters of this handbook, as they fit in the general synthesis we present here.

1.1. What is enterprise risk management

The following definition encompasses the currently accepted notions of ERM.

Definition 1.1 (*Enterprise risk management*). It is the strategy that aligns the firm's business with the risk factors of its environment in the pursuit of strategic objectives. It consists of the conceptual framework, organizational approaches, and tools that integrate market, credit, liquidity, operational, and business risks in achieving the organization's objectives.

We will adopt a specialized vocabulary that reflects applications of ERM to financial institutions, cognizant that enterprise risk management is applicable to other businesses as well. References to risk imply financial risk, which we define next.

Financial risk: It is the possibility that an unpredictable future event will result in a financial loss, with the consequence that an individual or institution will not meet some specified financial goals.

Financial risk may be characterize by the magnitude of the loss, its estimated likelihood today, and the causes of the event which are called *risk factors*. Risk is always in the future, as current or past losses do not present a risk. Also the future events must be unpredictable, but this does not restrict our ability to foresee plausible future events and plan for them. A hiker setting off on the trail can not predict the weather, but this does not restrict her from noticing that certain cloud formations could bring rain (plausible event) and carry foul weather gear.

The word "risk" derives from the old Italian word *risicare* which means "to dare". In this sense risk is more of a choice than a chance. Risk management is the discipline that provides tools to measure the risks and techniques to help us shape and make rational decisions about them. A survey of risk management techniques used by non-financial firms, published by Bodnar, Hayt and Marston (1998), reveals that more than half of the responding firms use derivatives in hedging their financial and other risks. The percentage exceeds more than 80% for the larger firms in the sample. Risk management is not restricted to financial institutions, and, furthermore, the distinction between financial and other forms of risk are becoming increasingly blurred, brought about by innovations such as weather derivatives, energy derivatives, catastrophe bonds and the like. The shift from a generic vocabulary to the specialized vocabulary focusing on financial institutions adopted in this chapter does not limit the applicability of ERM in the broader context.

We clarify now the concepts involved in the definition above.

Business: A financial institution's business is to provide services for (re)packaging and selling risks by *designing*, *pricing*, *capitalizing*, *funding*, and *marketing* financial products. Financial products are contractual agreements to pay customers contingent cashflows in the future, and they are funded by the assets of the institution.

Objectives: To use the cashflows generated by the business activities, and leverage debt or equity capital to enhance economic value through growth and consistency of earning, cashflow stability, and reduced costs of financial distress.

Market risks: These are the risks arising from changes in financial market prices and rates.

Credit risk: It is the risk of an unkept payment promise when an obligor—counterparty, issuer or borrower—defaults, or when their ability to make future payments comes into question and their quality rating changes.

Liquidity risks: These are the risks arising when an institution is unable to raise cash to fund its business activities (*funding liquidity*), or can not execute a transaction at the prevailing market prices due to a temporary lack of appetite for the transaction by other market players (*trading liquidity).*

Operational risks: These are the risks of potential losses due to human error, fraud, management failure, faulty systems or controls, corrupt data or erroneous models, and legal actions.

Business risks: These are the risks due to volatility of volumes, margins, or costs when engaging in the firm's business.

1.2. Example: Enterprise-wide view of credit risks in a bank

An example from banking provides a concrete background setting as we go through the general discussion. Credit risk is one of the central topics in risk management today. It is estimated that the total worldwide exposure in credit risky assets—loans, corporate bonds, and credit derivatives—exceeds 30 trillion USD. With a problem of this magnitude it comes as no surprise that substantial research has gone into the assessment of obligor creditworthiness, pricing of credit risky instruments, measurement and control of credit exposure, and analysis of portfolio credit losses.

For banks, in particular, credit risk appears both on the banking book and the trading book, covering various lines of business, diverse products, and customer markets. Figure 1 gives an enterprise-wide view of credit risk in banks. In some parts of the business, credit risk is a by-product of the market risk undertaken in the trading book by actively trading bonds, derivatives such as swaps and forwards, and credit derivatives. In other cases, such as the traditional lending businesses, it arises from actively originating, servicing and funding corporate loans or supporting retail-lending operations.

True ERM solutions for a bank are difficult to develop. The sources of credit risk as seen in Figure 1 are remarkably diverse. They come from small and medium business commercial loans, syndicated loans, retail mortgages and credit cards, corporate and sovereign bonds, credit derivatives, and so on. The expertise for managing these diverse risks was developed in different organizational units and credit risk management is traditionally managed strictly in silos. For instance, Commercial Banking manages mortgages, credit cards, and lines of credit, while the Treasury manages corporate bonds and credit derivatives. When risk is managed in silos the bank does not have an accurate view of its credit risk exposure to a specific borrower through the line of credit extended

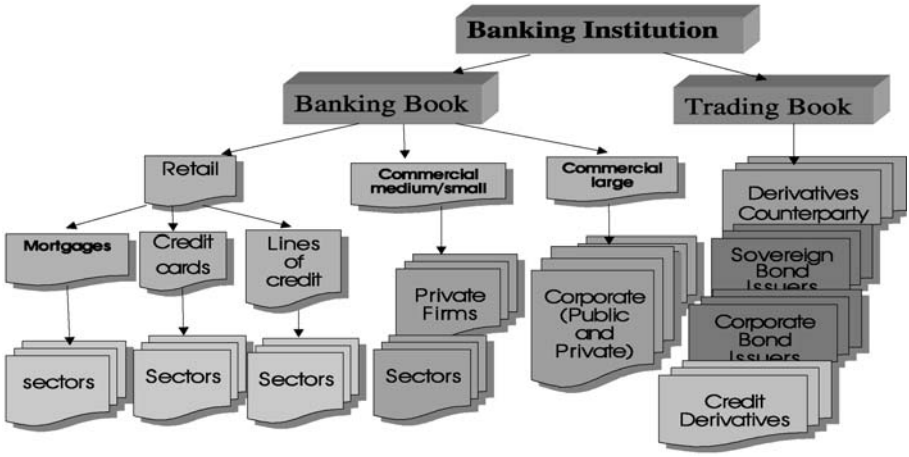


Fig. 1. Enterprise-wide view of credit risks in a bank.

to the borrower by Commercial Banking, and investments in its corporate bonds by the Treasury. Furthermore, credit risk has been traditionally managed separately from market risk, which has led to the emergence of different cultures within the bank. In today's integrated financial markets, managing in silos is not satisfactory and banks are compelled to manage the risks in their banking and trading books in a more unified manner adopting ERM.

2. A conceptual framework for enterprise risk management

The distinguishing features of ERM are the global view it takes of the enterprise, and an integrated view of the process. In order to achieve these features the components of an ERM system must be carefully planned. In this section we present a general framework for ERM and discuss its components.

In order to achieve the goal of aligning the firm's business with the risk factors of the environment, ERM must take a *global view* of the enterprise and the risks it is exposed to. Credit exposure to a single counterparty by a bank may appear through different lines of business, and a global view is needed if counterparty credit risks are to be identified. Similarly, country risk may appear through Treasury activities in a country's bond market or currency exchange contracts, but also through syndicated loans and lines of credit to corporations operating in the country, or derivative contracts underwritten by the country's banks. Again, a global view of the firm is needed to identify country specific risk.

Furthermore, in order to achieve the goal of alignment efficiently and add value, ERM must take an *integrated view* of the business processes. As *concurrent engineering* calls for the integration of engineering design, manufacturing, and marketing of products

in a seamless process managed by a product team (Hauser and Clausing, 1988), similarly, *integrated financial product management* calls for the integration of the functions involved in designing, pricing, capitalizing, marketing, and funding financial products (Holmer and Zenios, 1995). These functions are clearly interdependent. The design of a financial product will affect its price and the asset allocation decisions for funding it, as the design of an automobile affects the manufacturing process and, eventually, its cost. The interdependencies, and their effects on the return and risk profile of a product, can only be analyzed when taking an integrated view of these functions as parts of a process.

When multiple financial products are offered by an institution there is the additional problem of managing the business portfolio. Determining the appropriate product mix and allocating the firm's capital should again take an integrated view of the risks and returns of competing lines of business.

We consider in more detail the functions that must be integrated in a process, taking first the micro-management perspective (i.e., for a single line of business) and then the macro-management perspective (i.e., for a business portfolio).

2.1. The management of a single line of business

The business of a financial institution is to re-package financial risks, to price them, and either to transfer them to other market participants or hold them in a well-diversified portfolio. The means for engaging in this business is the sale of financial products. We consider here an institution that offers a single product. The sale of the product creates contractual obligations to pay customers cashflows in the future. The payments may be contingent on some underlying risk factors, such as an accident—in the case of insurance products—or a market movement—in the case of derivative securities—or snowfall—in the case of weather derivatives. The firm's financial obligations may stretch well into the future and are usually uncertain.

In order to align this single line of business with the risk factors of the environment, management engages in the following functions: *designing* the product, *pricing* the product in a way that is consistent with its risk profile and the prices of other instruments in the financial markets, and making decisions on asset allocation and asset management to *fund* the product. These functions correspond to the repackaging of the risks, pricing them and transferring or diversifying them. These functions are part of the tactical management process, see Figure 2, and should be managed in an integrated way.

The three functions are coordinated so that the risk adjusted return from the process is maximized. The return calculations require prices, while risk adjustments require probability estimates and risk preferences. Lo (1999) discusses the three P's of ERM: probabilities of extreme events, asset prices, and preferences towards risk. The P's are applicable to all three functions—designing, pricing, and funding—and should be studied for the integrated process, and not for each function separately. We look at each one of the three functions and discuss how the P's apply.

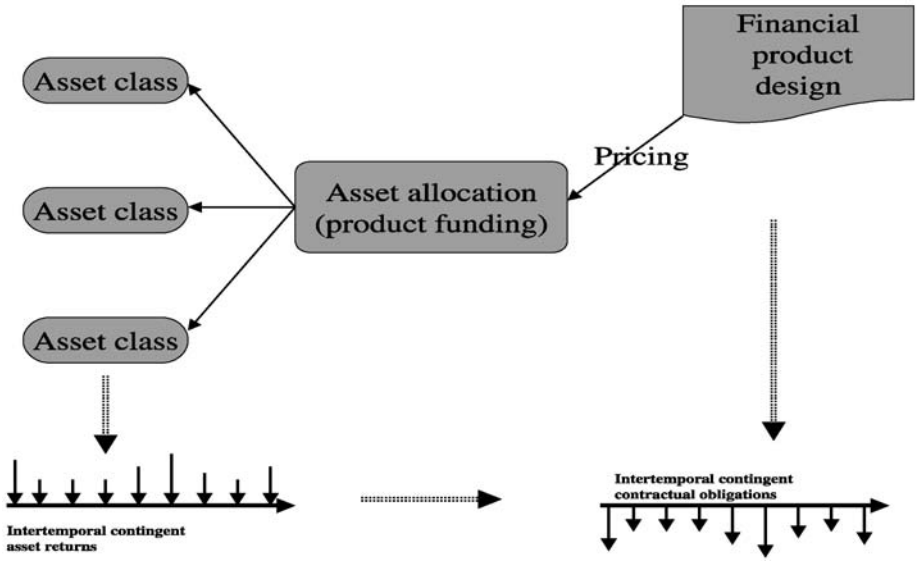


Fig. 2. The functions of a firm offering a single product (boxes) and their interdependence into an integrated process (arrows).

2.1.1. Designing the product

Modern financial products are very complex. They possess multiple distinct attributes that may be attractive to different investors, but will also affect the price of the product. Consider for instance participating insurance policies with guarantees (see the chapter by Consiglio, Cocco and Zenios in the second volume of this handbook). This product is characterized by the minimum guaranteed rate of return, the participation rate in the asset portfolio returns with the disbursement of bonuses, and surrender charges. High guaranteed returns are attractive to investors but the product may be expensive to fund. Excessively high guarantees may even lead to bankruptcy, as was the case with Nissan Mutual Life that failed on a $2.56 billion liability arising from a 4.7% guaranteed investment. Reducing the guaranteed rate and increasing the participation rate makes the product less attractive but easier to fund.

The design of financial products remains an art, although substantial research has gone into explaining financial innovation as the means for risk sharing (Allen and Gale, 1994). For an example on the scientific approach for designing financial products for the Federal National Mortgage Association see Consiglio and Zenios (1997, 1999). The probabilities associated with the risk factors underlying the product must be assessed, together with the buyers' preferences, in order to arrive at a fair structure for product and its associated price. Marketing considerations are also important in the design of financial products although not much has been done in this area either (Wind, 1987).

2.1.2. Pricing the product

Asset pricing is one of the great successes of modern financial economics. The ideas of Black and Scholes (1973) and Merton (1973) have spread with remarkable speed in both academia and industry. To the heart of the function of asset pricing is the replication of complex securities by sophisticated arbitrage-free strategies involving considerably simpler instruments. Probabilities of changes in the risk factors underlying the priced asset are essential in structuring the replicating portfolio. Therefore, pricing new products is not independent of the design specifications discussed above. Risk preferences are included through implicit assumptions when dealing with pricing in complete markets using replicating strategies (Lo, 1999), and they are needed explicitly for pricing in incomplete markets, such as in evaluating insurance liabilities or pension fund claims (Dembo, Mausser and Saunders, 2000).

2.1.3. Funding the product

Funding the financial product entails the design of an asset portfolio whose cashflow pattern will replicate the contractual obligations of the product into the future and under different contingencies. Probabilities of the risk factors, that drive price changes for both the contractual obligations and the asset portfolio movements, are important. Price changes of the contractual obligations and the assets must be managed within a portfolio strategy that closely replicates obligations both across time and across states of the risk factors. Close tracking of the obligations may of course be expensive, and the institution's risk preferences guide the selection of a portfolio trading off tracking error against excess returns.

2.2. The management of a business portfolio

Financial institutions typically support multiple lines of business. In addition to resolving the tactical issues relating to the design, pricing, and funding of each line of business, the firm must deal with the strategic management of the business portfolio of multiple and diverse products. The firm's balance sheet can be viewed as a portfolio composed of assets—representing investment portfolios—and liabilities—representing specific products. The difference between asset and liability values is the firm's equity. From a macro-perspective the focus is to allocate equity capital among the competing lines of business in a way that is consistent with the risks and potential rewards of the products offered by each one. Maximizing the risk-adjusted return on equity is the strategic objective in this case, which must be tempered by the need to satisfy risk-based capitalization rules and other regulatory requirements.

2.3. Integrating design, pricing, funding, and capitalization

The integration of the design, pricing, and funding of each product results into a process that maximizes risk adjusted return for each product. The risk adjusted return of a

product, however, is not independent from the composition of the overall portfolio, an observation first made by Markowitz (1952) with profound implications for the theory of portfolio choice (see also Chapter 2 in this volume). In particular if the risks of a product are not perfectly correlated with other risk factors in the portfolio the overall risk exposure of the portfolio can be reduced. The allocation of equity capital based on the risk adjusted return of each product must recognize that some lines of business act as natural hedges against each other. This can be achieved only if the management of the business portfolio is integrated with the management of the distinct processes at an enterprise level, see Figure 3. The functions of designing, pricing, and funding a product are integrated into a process (see Figure 2) and capital allocation is determined among these competing processes (Figure 3). The tactical decisions for the management of a single product provide only marginal estimates of the risk adjusted return. However, consistency in the management of single lines of business by identifying the underlying risk factors and modelling them in a common framework, permits ERM to take a portfolio view when integrating the design, pricing, funding and capitalization of multiple products.

This analysis has identified the levers that management can pull in aligning its business with the risk factors of the environment. These are the decisions on the design and pricing of a product, allocation of assets in funding the product, and equity allocation in capitalizing the firm's lines of business. These levers are effective in aligning risk factors only when management takes a global view of the enterprise, and an integrated view of the processes involved. The traditional approach of managing risks in silos is not effective because it gives neither a view of the processes, nor a global view of the enterprise.

But how are these levers to be managed? How is the design, pricing, funding and equity allocation for products to be executed? The general framework has yet to offer concrete tools to answer these questions. The essential components that provide management with the tools to support the general ERM framework are discussed next.

2.4. Components of enterprise risk management

Neither academic research nor the experience of end-users and solution providers has produced a definitive description of the components comprising the ERM framework. See, for instance, Lam (1999a, 1999b), the books by Stulz (2001), Crouhy, Galai and Mark (2001) or the handbook by Lore and Borodovsky (2000).

The components of effective ERM are the following:

1. Risk measurement,
2. Risk management,
3. Performance measurement, and
4. Corporate governance.

These components provide a broad description of the components of ERM systems in which currently suggested approaches would fit. We describe the tasks involved in

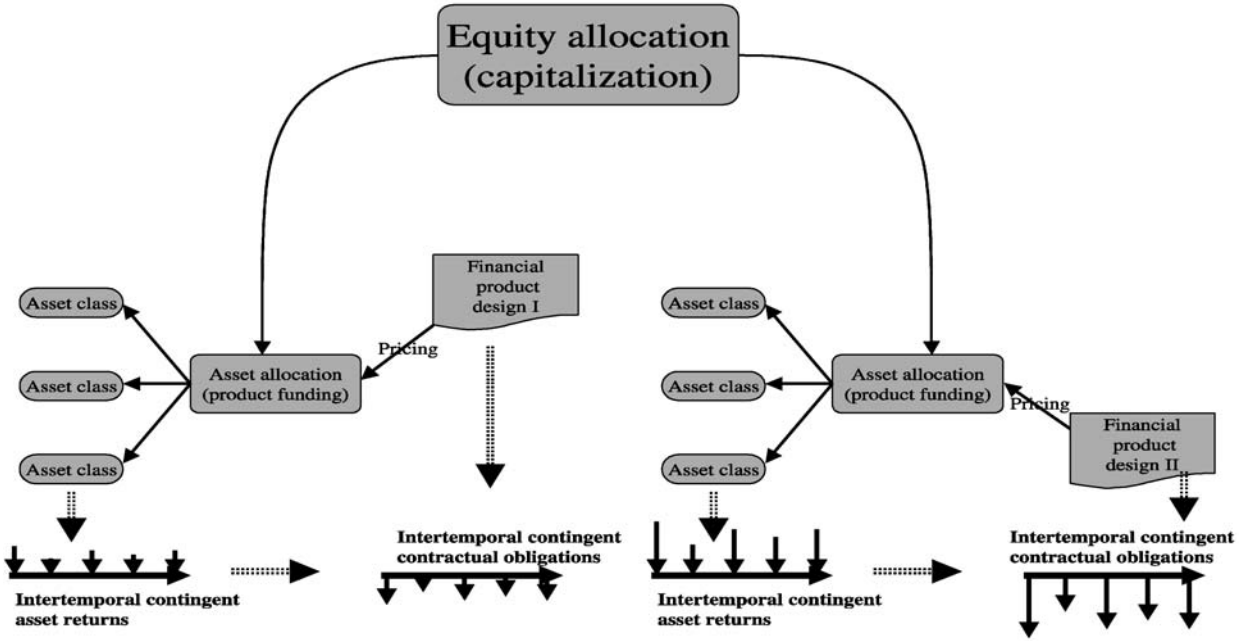


Fig. 3. Managing a portfolio of products offered by multiple lines of business and their interdependence through integrated processes.

each one of the four components. However, details on the organizational approaches, and the tools needed to achieve these tasks are discussed only in passing.

2.4.1. Risk measurement

Risk measurement first identifies the risk factors that arise in the operation of the business and in the environment. Some risks are connected to the core activities of the firm (core exposures) and some to the facilitating activities (peripheral exposures). For instance, actuarial risk is core exposure for an insurer, while credit risk is peripheral exposure in managing the insurer's asset portfolio. For a Bank, however, credit risk is the core exposure of its banking book.

Risk measurement analyzes how these separate risks interact at the enterprise level and how they influence business performance. The end product of this analysis is a distribution of profits and losses under different future realizations of the risk factors, see Figure 4. From this distribution various risk metrics can be computed. For instance, the classical mean-variance portfolio theory uses the variance of the distribution as a measure of risk, see Chapter 2. Value-at-Risk (VaR) estimates a left quantile as illustrated in the figure. These calculations have become an industry standard for risk measurement especially in banking and asset management, see, for example, Jorion (1996). Regulators also rely on them, see the Basel Committee on Banking Supervision (1996a, 1996b, 2001) reports. They are used for assigning risk capital and monitoring risk exposures.

However, VaR suffers from a severe shortcoming that limits its use in supporting ERM. In particular it is not sub-additive, that is the sum of the VaR of two positions may be less than the VaR of a portfolio of these positions. Thus while it may appear that the total risk exposure, measured by adding up the individual risk exposures is within acceptable limits, the actual risk exposure of the total position may be unacceptably high. Put another way, by splitting a portfolio into multiple positions one may estimate a much lower risk exposure than what is captured by the VaR of the original portfolio.

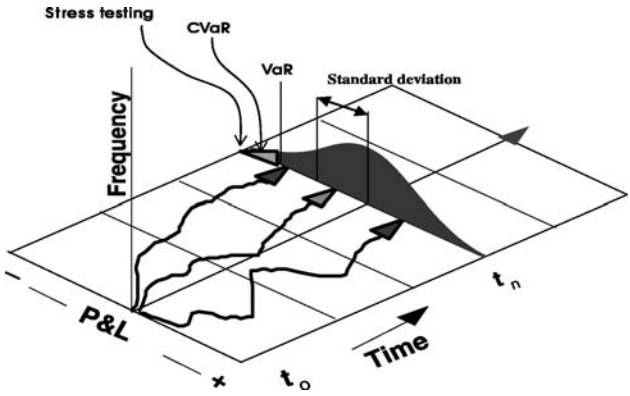


Fig. 4. Distribution of profits and losses and the estimation of various risk measures.

Sub-additivity ensures that the risk measure is reasonable when adding two positions, and it allows the decentralized calculation of the risk at an enterprise wide level. In particular the sum of the risks of individual positions provides an upper bound on the enterprise wide risk. Axiomatic characterizations of risk measures that can be used to support ERM exist in the form of *coherent* risk measures (Artzner et al., 1999). Coherent risk measures such as conditional Value-at-Risk (CVaR) or expected shortfall have a long tradition in the insurance industry, Embrechts, Klüppelberg and Mikosch (2000) and the work by Cariño and Ziemba (1998) and Cariño et al. (1994, 1998) on the Japanese insurance industry. Such models have been gaining popularity recently in asset management as well (Rockafellar and Uryasev, 2000). They have been proven particularly effective in managing credit risk (Andersson et al., 2001; Jobst and Zenios, 2001). Several other measures of risk exist in the literature, satisfy the coherence axioms and find different levels of acceptance by end-users. We mention, for instance, mean absolute deviation and its variants (Sharpe, 1971; Konno and Yamazaki, 1991; Zenios and Kang, 1993) and regret (Dembo and King, 1992; Dembo and Rosen, 1999; Mausser and Rosen, 1999).

For large and complex positions the calculation of VaR or CVaR requires Monte Carlo simulations of the underlying risk factors, and repricing the positions with changes of the risk factors. When estimating risk exposures for ERM we need to recognize the multiple interrelated risk factors, and their effect on the firm's portfolio. The dynamics of the risk profile should also be recognized (Mulvey and Zenios, 1994b).

Risk analytics support the quantification of the disparate sources of risk such as market, credit, liquidity, and operational risk. Analytics are also used to quantify business risk, such as, for instance, the contractual obligations to pension fund participants or insurance policyholders, volume changes due to policy surrenders and the like. Dembo et al. (2000) present a general framework for analytics to support the risk measurement tasks of ERM.

2.4.2. Risk management

The risk profiles identified through risk measurement must be shaped into an overall risk profile for the institution. The overall risk profile does not just happen as the sum of the risk profiles of individual lines of business. Instead it should be shaped to achieve the firm's objectives, while satisfying regulatory restrictions and business policy requirements. In the most basic approach the risk profile can be shaped by controlling the risk of the individual positions through limits management. However enterprise risk management goes beyond *risk control*. It also recognizes diversification benefits and natural hedges, and uses the financial markets to effectively and efficiently re-allocate specific risks.

In order to support diversification, hedging and risk transfer a risk management system must provide the following functionality:

1. *Portfolio compression* to represent complex portfolios simply.
2. *Risk decomposition* to decompose risk by asset and/or risk factors and quantify the contribution of each asset or risk factor to an overall portfolio.
3. "*What if*" *analysis* to understand how new trades affect the portfolio risk.
4. *Hedging and optimization* of complex portfolios.

To these core functionality we may add systems to understand the impact of instruments' non-linearities and of non-normal risk factor distributions on portfolio risks, and to understand complex, non-intuitive, market views implicit in the portfolio as well as in the investment policy or market liquidity (implied views).

Analytics that support the decomposition of the risk measures of the previous section so that risk management can be executed require approximations of the portfolio sensitivity to the risk factors (Litterman, 1989; Garman, 1996, 1997). Simulation-based tools (Mulvey and Zenios, 1994a, 1994b; Mausser and Rosen, 1998, 1999; Jobst and Zenios, 2001) provide additional insights when the portfolio contains non-linearities, when the market distributions are not normal or when there are multiple horizons. In particular, simulations are useful not only for market risk, but also for credit risk, where the exposure and loss distributions are generally skewed and far from normal. Furthermore simulation-based tools can be used, sometimes even more efficiently, with other risk measures in addition to VaR such as CVaR or expected regret.

Risk transfer strategies are used to lower the cost of undesirable risks and to increase the firm's capacity to originate risks in its core exposure. In the course of doing business institutions develop risk concentrations in their area of specialization. Risk transfer of the portfolio to the secondary markets increases the capacity for risk origination without accumulating highly concentrated risk positions. Lam (1999b) reports savings of the order of 20 to 30% in the cost of risk transfer by institutions taking an enterprise wide approach, bundling together insurance protection with financial risk protection. Derivative securities and insurance products together with well diversified portfolios of products lower the costs of financial distress (Stulz, 1996).

Finally allocation of equity capital, among competing lines of business taking into account the risk and return profiles of each, creates well diversified portfolios with reduced risk exposure. Integrated financial product management plays a key role here. The task of risk management is completed with the use of portfolio optimization models (Zenios, 1993; Ziemba and Mulvey, 1998; Zenios, 2006).

2.4.3. Performance measurement

Having estimated the risk profile of different lines of business and shaped the risk profile of the enterprise, ERM turns to the problem of performance measurement. The task now is to estimate the contribution of each risk factor to the firm's diversified overall risk profile. This is the feedback loop in the ERM process. Performance measurement serves in the decentralized management of the enterprise by setting goals for line managers when risks are first acquired, determining local hedges and capital equity allocation, and providing reward incentives. Performance measurement derived from the steps of

risk measurement and risk management outlined earlier ensures firm-wide consistency in the handling of risks.

Stakeholder management also requires risk transparency that is achieved through a firm-wide performance measurement system. Regulators need to know that business operations are in compliance with regulatory requirements. Equity analysts and rating agencies need risk information to develop investment and credit opinions. Unless there is both external and internal risk transparency the institution may be penalized for its risks as the stakeholders may see only the risks but not the controls, or the natural hedges that may exist among them. Performance measurement ensures that the required transparency is achieved. This is also of use in the analysis of mergers and acquisitions.

2.4.4. Corporate governance

None of the components of ERM as outlined above will achieve its tasks without the support of appropriate corporate governance. Guidelines here are scant. For instance, the handbook on financial risk management endorsed by the Global Association of Risk Professionals (Lore and Borodovsky, 2000) does not contain any chapters on the topic of corporate governance for ERM. Some discussion on this topic is given in Lam (1999b). Santomero (1997) and Santomero and Babbel (1997) analyze the process of risk management in Banks and Insurance firms.

Corporate management ensures that the components of the ERM are in place, they function properly and are aligned with each other and with the objective of the enterprise risk management strategy. To do so management must define the institution's risk appetite in terms of loss tolerance, risk-to-capital leverage, and target debt rating. Risk management skills must be available to support the ERM strategy in the form of both human, technological, and financial resources. An organizational structure that allows the enterprise wide management of risk should clearly define the roles of risk management at both the enterprise level and by the line managers. Figure 5 illustrates the organizational structure of a bank with the addition of a Chief Risk Officer (CRO). The Tillinghast-Towers Perrin finds that only 20% of the respondents have appointed a CRO, with a noticeable difference between North American institutions (8% have appointed CRO) and institutions in the rest of the world (40% have appointed CRO). The survey concludes, based on experience inside and outside the financial services industry, that creating a CRO function can be an effective step in making ERM work in practice.

We concluded the section describing the conceptual framework for ERM. Before we introduce asset and liability management as a core activity of ERM we ask what is it that we achieve by taking a global and integrated view of risk management.

2.5. Why is enterprise risk management important

Financial risks come in many forms and quite often are interdependent. ERM allows an institution to identify the many and disparate sources of risk in its environment, measure their interdependence and exploit natural hedges that appear in its line of business. This

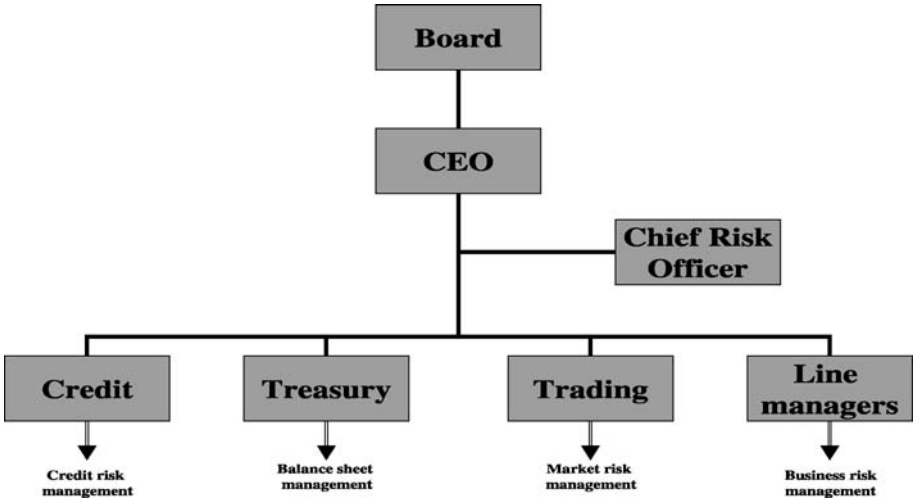


Fig. 5. Organization structure to support enterprise risk management.

leads to more efficient capital allocation. Integration also leads to the design of innovative risk transfer products. Recall that according to Lam (1999b) bundling together insurance protection with financial risk resulted in savings of the order of 20 to 30% in the cost of risk transfer. ERM also allows the simultaneous management of the business risks and those of the trading book. This allows for more efficient use of the financial markets in transferring concentrated business risk, a practice that has been used with success in re-insurance, weather derivatives, and energy contracts. ERM also allows for more efficient risk transferring internally, since transfer pricing becomes less of a political process among the silos and can be viewed at an integrated enterprise level in conjunction with the firm's external risk factors.

In the domain of credit risk ERM facilitates the credit risk transfer pricing between loan origination and portfolio management. The cost of distress from credit losses and credit value volatility is reduced, and liquidity of credit risk instruments is improved, leading to better trading and hedging decisions. Credit risk capital can be allocated more efficiently since various business opportunities are directly comparable.

Overall ERM leads to greater risk transparency for both shareholders and regulators. Indeed, some of the motivation for introducing ERM comes from external pressures from corporate governance bodies and institutional investors. We give some examples. The BIS 1988 accord allows banks to use internal models for determining minimum regulatory capital for market risk of the traded positions—see the report by the Basel Committee on Banking Supervision (1996a, 1996b). This led to a proliferation of risk measurement systems in Banks. The New BIS Capital Accord emerging out of the consultative papers issued by Basel Committee on Banking Supervision (2001) extends the use of approved internal models to consider credit risk, market risk for non-traded

positions, and operational risk. The three pillars of sound risk management practices according to the new accord are regulatory requirements for capital, supervisory review processes, and the market discipline imposed by shareholder pressures when risks are properly disclosed. Risk transparency is key to all three pillars and the significance of ERM increases within the new BIS accord.

3. Asset and liability management in enterprise risk management

The management of the firm's balance sheet is at the core of ERM for financial institutions. For non-financial institutions balance sheet management is traditionally viewed as solely an accounting issue, although this is also changing (Bodnar, Hayt and Marston, 1998). For financial institutions, however, the balance sheet reflects the risks of the environment from the asset side and most of the business risks from the liability side. Aligning these risks is the goal of an asset and liability management system. ALM addresses parts of the problems of ERM. As such it provides tools for risk measurement and risk management that are applicable to ERM. ALM has a history that dates back to Markowitz seminal contributions in the 1950s (Markowitz, 1952, 1991) for asset allocation, and the more recent extensions to include liabilities (Sharpe and Tint, 1990). Tools for ALM are better developed than those for ERM although they may not be as advanced as needed to support ERM, see the volume published by the Kamakura Corporation (1998). In order for ALM to be effective in supporting the broader goals of ERM is must receive additional information. The interaction between ERM and ALM are discussed here.

ALM takes a more focused view of risks than ERM. It focuses on market, credit, and liquidity risk on the asset side. On the liability side it focuses on volatilities of margins and costs. Volatility of sales volumes is not directly part of the ALM function although it is a significant component of business risk. Integrating volatility of sales volumes requires aligning with the marketing environment. Similarly, the management of operational risk requires an alignment of the business with the operational environment, and falls outside the scope of ALM in the domain of ERM.

Managing the financial risks of the balance sheet can be achieved by pulling the levers of ERM. In particular asset allocation (funding the products) and equity allocation (capitalization of the products) are the major managerial activities in ALM. The design and pricing of products are usually considered as exogenously given in the context of ALM, although pricing is used as input to the asset allocation phase and in determining hedging positions.

3.1. Components of asset and liability management

The basic components that make up an ALM system are the following:
1. Data storage,
2. Analysis tools, and
3. Reporting facilities.

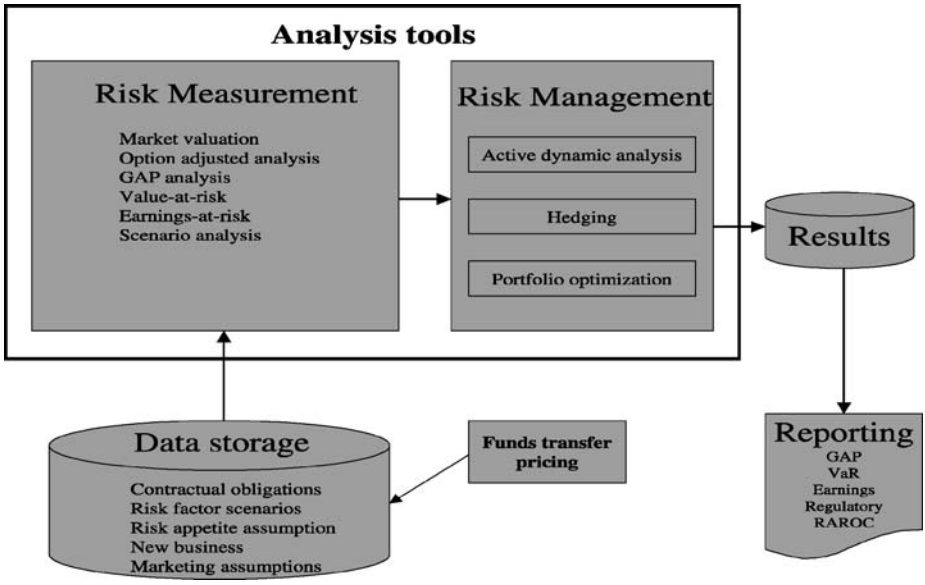


Fig. 6. Components of an asset and liability management system.

The ALM system is illustrated in Figure 6.

The goals of the system and its components are to perform tasks such as earnings and balance sheet simulation, sensitivity analysis, current and future valuation, and dynamic balance sheet modelling. These functions provide the necessary input data for management to align the asset and liability sides of the balance sheet through asset allocation or equity capital allocation. Aligning the assets and liability sides is achieved when the net risk position is neutralized. For instance, duration matching can be used to neutralize market risk due to parallel shifts of the term structure.

Data storage provides data on the contractual obligations, market information such as prices and probabilities of changes in the risk factors, behavioral assumptions about risk appetite, and information about business volumes, customer attrition and new products. Historical data may be necessary in calibrating other model parameters as well and they are also provided through data storage.

Reporting provides the necessary information to regulators and shareholders, such as earnings at risk, value at risk, capital adequacy requirement and so on. It also serves to answer "what if" questions to support managerial decisions.

Analysis provides the analytic support for both risk measurement and risk management. Analytics are used to support the quantification of the disparate sources of risk, and facilitate the use of derivative securities and portfolio optimization models to neutralize these risks. ALM analytics are a multilayer application. At the base level there are some mechanical features such as cashflow generators, gap analysis and duration and convexity mismatch calculations. Next comes the analysis of the economic en-

vironment. This typically takes the form of forecasts or simulations of interest rates, exchange rates, and credit spreads. The projected estimates or the scenarios are used to simulate the balance sheet by repricing the assets and liabilities under the assumed scenarios. Dembo et al. (2000) provide a general framework to support this analysis, and the chapters in Part I of this handbook cover a broad spectrum of asset and liability modelling techniques.

Valuation functionality is an important component of ALM analytics, not only for the rolling horizon simulations but also for pricing what is on the balance sheet today. For some items on the balance sheet this may be straightforward. For products with embedded options analytics are needed to capture the highly asymmetric value profiles. Collateralized Debt Obligations for banks, or participating policies with guarantees for insurers are examples of products whose valuation requires mathematical sophistication.

The simulations can be extended well into the future and the institution's balance sheet can be simulated dynamically across time (Mulvey and Zenios, 1994b). Dynamic analysis can be either passive, focusing on the time-evolution of the current balance sheet. Or it can be active, and focus on the analysis of explicit institutional ALM strategies (Perold and Sharpe, 1988).

4. Models for asset and liability management

We conclude this chapter with a classification of ALM models. Details on various models are given in the subsequent chapters of this handbook. A general classification is possible based on how they address risk management problems with respect to time and uncertainty.

The time-axis may consist of a single-period from today to the end of the horizon. This view is static since no changes are possible after today and before the end of the horizon. Static time models assume that there is only today, $t = 0$, and only one portfolio decision is made. Decisions to be made in the future are not explicitly incorporated in the optimization model. The time-axis may also be multi-period taking a dynamic view of the ALM problems. Dynamic time models recognize that portfolio decisions will be made at future time periods, $t = 1, 2, \ldots, T$, and these decisions are explicitly modelled. This is a more realistic setting since scenarios are observed and actions are possible at several discrete points in time from today until the end of the horizon.

Similarly we may view the risk factors as being static or stochastic. Under a static assumption the risk factors—the economy, asset returns, the term structure of interest rates—will remain at their current state and will at most change with only small shifts. Stochastic models assume that risk factors evolve with time according to some probability distribution. Scenarios drawn from this distribution are explicitly incorporated in the model.

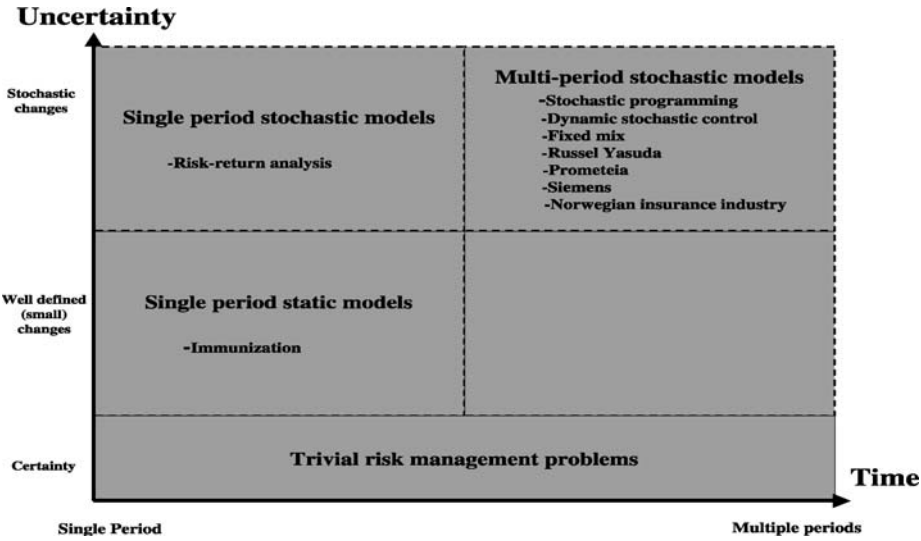


Fig. 7. Classification of portfolio optimization models in the dimensions of time and uncertainty, and the models of this handbook.

Figure 7 illustrates a classification of ALM models along the dimensions of time and uncertainty, and places the models of this Handbook in the appropriate class.

Single-period static models: Such models hedge against small changes from the current state of the world. For example, a term structure is input to the model which matches assets and liabilities under this structure. Conditions are then imposed to guarantee that if the term structure shifts somewhat from the current state, the assets and liabilities will move in the same direction and by equal amounts. This is the fundamental principle behind *portfolio immunization* introduced by Reddington (1952), see de La Grandville (2001) for an up-to-date treatment. A static model does not permit the specification of a random variable to describe changes of the economic environment, something that is achieved with the models of the next class.

Single-period stochastic models: A stochastic model describes the *distribution* of returns of both assets and liabilities in the volatile environment, and ensures that movements of both sides of the balance sheet are highly correlated. This is the idea pioneered by Markowitz (1952, 1991) with the mean-variance models in his seminal papers for asset allocation, and extended throughout the years to include liabilities (Sharpe and Tint, 1990, and the volume by Kamakura Corporation, 1998) and asymmetrically distributed non-normal returns (Mossin, 1987; Hakansson, 1974; Zenios, 1995).

A single period stochastic model is *myopic*. That is, it builds a portfolio that will have a well behaved distribution of the gap between asset and liability returns under the assumed scenarios. However, it does not account for the fact that the portfo-

lio manager is likely to rebalance the portfolio once some surplus is realized or in response to shortfall. Furthermore, as the scenarios evolve across time different portfolios may be more appropriate for capturing the correlations of assets and liabilities. The single-period model may recommend a conservative strategy, while a more aggressive approach may be justified once we explicitly recognize the manager's ability to rebalance the portfolio. While it has been established that under some conditions myopic policies are optimal in a dynamic setting (Mossin, 1987; Hakansson, 1974) they remain restrictive for several important problem classes, especially when there are exogenous liabilities with long time horizons. The limitations of a myopic model are overcome by the multiperiod stochastic models discussed below.

Multiperiod static models: A multiperiod static model will explicitly model portfolio rebalancing decisions at future time periods, but it will assume that the state of the world will shift with small and well-defined changes during this period. A model based on these assumptions is equivalent to a decision maker with perfect foresight. Ziemba and Vickson (1998, Section IV) deal with some important papers in this category. In general, however, this is an unrealistic setting, and multiperiod models that allow for rebalancing in a static world are of limited practical use.

Multiperiod stochastic models: Multiperiod stochastic models allow both assets and liabilities to evolve dynamically across time following a probability distribution. Portfolio decisions are revised as time evolves and more information is received about the assets and the liabilities. Most financial decisions are not irreversible, and their reversibility is an option that must be explicitly incorporated in the model. Multi-period stochastic models capture both the stochastic nature of the problem and the fact that the portfolio decisions are reversible. Mathematical models under the general term of *stochastic programming with recourse* provide the framework for dealing with this broad problem and modelling dynamic strategies (Ziemba and Vickson, 1975, 2006; Zenios, 1993; Birge and Louveaux, 1997; Ziemba and Mulvey, 1998).

References

Allen, F., Gale, D., 1994. Financial Innovation and Risk Sharing. The MIT Press, Cambridge, MA.

Andersson, F., Mausser, H., Rosen, D., Uryasev, S., 2001. Credit risk optimization with conditional Value-at-Risk criterion. Mathematical Programming, Series B 89, 273–291.

Andrews, K.R., 1971. The Concept of Corporate Strategy. Dow-Jones Irwin, New York.

Artzner, P., Delbaen, F., Eber, J.-M., Heath, D., 1999. Coherent measures of risk. Mathematical Finance 9, 203–228.

Basel Committee on Banking Supervision, 1996a. Overview of the amendment to the capital accord to incorporate market risks. Basel committee publications no. 23, Bank for International Settlement, Basel, Switzerland. http://www.bis.org/.

Basel Committee on Banking Supervision, 1996b. Amendment to the capital accord to incorporate market risks. Basel committee publications no. 24, Bank for International Settlement, Basel, Switzerland. http://www.bis.org/. Modified in September 1997.

Basel Committee on Banking Supervision, 2001. The new Basel capital accord. Consultative document, Bank for International Settlement, Basel, Switzerland, January.

Birge, J.R., Louveaux, F., 1997. Introduction to Stochastic Programming. Springer-Verlag, Heidelberg.
Black, F., Scholes, M.J., 1973. The pricing of options and corporate liabilities. Journal of Political Economy 81, 637–654.
Bodnar, G.M., Hayt, G.S., Marston, R.C., 1998. 1998 Wharton survey of financial risk management by US non-financial firms. Financial Management 27 (4), 70–91.
Cariño, D.R., Myers, D.H., Ziemba, W.T., 1998. Concepts, technical issues, and uses of the Russell–Yasuda Kasai financial planning model. Operations Research 46, 450–462.
Cariño, D.R., Ziemba, W.T., 1998. Formulation of the Russell–Yasuda Kasai financial planning model. Operations Research 46, 433–449.
Cariño, D.R., Kent, T., Myers, D.H., Stacy, C., Sylvanus, M., Turner, A.L., Watanabe, K., Ziemba, W.T., 1994. The Russell–Yasuda Kasai model: An asset/liability model for a Japanese insurance company using multistage stochastic programming. Interfaces 24 (1), 29–49.
Consiglio, A., Zenios, S.A., 1997. Optimal design of callable bonds using tabu search. Journal of Economic Dynamics and Control 21, 1445–1470.
Consiglio, A., Zenios, S.A., 1999. Designing portfolios of financial products via integrated simulation and optimization models. Operations Research 47, 195–208.
Crouhy, M., Galai, D., Mark, R., 2001. Risk Management. McGraw-Hill, New York.
de La Grandville, O., 2001. Bond Pricing and Portfolio Analysis. The MIT Press, Cambridge, MA.
Dembo, R.S., King, A.J., 1992. Tracking models and the optimal regret distribution in asset allocation. Applied Stochastic Models and Data Analysis 8, 151–157.
Dembo, R., Mausser, H., Saunders, D., 2000. Valuation in incomplete markets: An optimization approach. Algo Research Quarterly 3 (2), 29–37.
Dembo, R., Rosen, D., 1999. The practice of portfolio replication. Annals of Operations Research 85, 267–284.
Dembo, R.S., Aziz, A.R., Rosen, D., Zerbs, M., 2000. Mark to Future: A Framework for Measuring Risk and Reward. Algorithmics Publications. Algorithmics Inc., Toronto, May.
Embrechts, P., Klüppelberg, C., Mikosch, T., 2000. Modelling Extremal Events for Insurance and Finance. Applications of Mathematics, Stochastic Modeling and Applied Probability, vol. 33. Springer-Verlag, Heidelberg.
Garman, M., 1996. Improving on VaR. RISK 9 (5), 61–63.
Garman, M., 1997. Taking VaR to pieces. RISK 10 (10), 70–71.
Hakansson, N.H., 1974. Convergence to isoelastic utility and policy in multiperiod portfolio choice. Journal of Financial Economics, 201–224.
Harker, P.T., Zenios, S.A. (Eds.), 2000. Performance of Financial Institutions: Efficiency, Innovation, Regulations. Cambridge Univ. Press, Cambridge, England.
Hauser, J.R., Clausing, D., 1988. The house of quality. Harvard Business Review, 63–73, May–June.
Holmer, M.R., Zenios, S.A., 1995. The productivity of financial intermediation and the technology of financial product management. Operations Research 43 (6), 970–982.
Jobst, N.J., Zenios, S.A., 2001. The tail that wags the dog: Integrating credit risk in asset portfolios. Journal of Risk Finance, 31–43, Fall.
Jorion, P., 1996. Value at Risk: The New Benchmark for Controlling Market Risk. Irwin Professional Publishing, IL.
Kamakura Corporation, 1998. Asset and Liability Management: A Synthesis of New Methodologies. Risk Publications.
Konno, H., Yamazaki, H., 1991. Mean absolute deviation portfolio optimization model and its applications to Tokyo stock market. Management Science 37 (5), 519–531.
Lam, J.C., 1999a. Enterprise-wide risk management: staying ahead of the convergence curve. Journal of Lending & Credit Risk Management 81 (10), 16–20.
Lam, J.C., 1999b. Integrated risk management. In: Derivative Credit Risk: Further Advances in Measurement and Management. Risk Books.
Litterman, R., 1989. Hot spots and hedges. Journal of Portfolio Management, 49–54, Summer.

Lo, A.W., 1999. The three P's of total risk management. Financial Analysts Journal, 51–57, January/February.
Lore, M., Borodovsky, L. (Eds.), 2000. The Professional's Handbook of Financial Risk Management. Butterworth Heinemann, Oxford, England.
Markowitz, H.M., 1952. Portfolio selection. Journal of Finance 7, 77–91.
Markowitz, H.M., 1991. Portfolio Selection: Efficient Diversification of Investments, second ed. Basil Blackwell, Oxford.
Mausser, H., Rosen, D., 1998. Beyond VaR: from measuring risk to managing risk. Algo Research Quarterly 1 (2), 5–20.
Mausser, H., Rosen, D., 1999. Beyond VaR: triangular risk decomposition. Algo Research Quarterly 2 (1), 31–43.
Merton, R.C., 1973. Theory of rational option pricing. Bell Journal of Economics and Management Science 4, 141–183.
Miccolis, J., Shah, S., 2000. Enterprise risk management: An analytic approach. Monograph, Tillinghast–Towers Perrin, New York.
Mossin, J., 1987. Optimal multiperiod portfolio policies. Journal of Business 41, 215–229.
Mulvey, J.M., Zenios, S.A., 1994a. Capturing the correlations of fixed-income instruments. Management Science 40, 1329–1342.
Mulvey, J.M., Zenios, S.A., 1994b. Diversifying fixed-income portfolios: modeling dynamic effects. Financial Analysts Journal, 30–38, January/February.
Perold, A.F., Sharpe, W.F., 1988. Dynamic strategies for asset allocation. Financial Analysts Journal, 16–27, January–February.
Porter, M.E. (Ed.), 1980. Competitive Strategy. Free Press, New York.
Reddington, F.M., 1952. Review of the principles of life-office valuations. Journal of the Institute of Actuaries 78, 286–340.
Rockafellar, R.T., Uryasev, S., 2000. Optimization of conditional Value-at-Risk. The Journal of Risk 2 (3), 21–41.
Santomero, A.M., 1997. Commercial bank risk management: an analysis of the process. Journal of Financial Services Research.
Santomero, A.M., Babbel, D.F., 1997. Financial risk management by insurers: an analysis of the process. The Journal of Risk and Insurance 64, 231–270.
Sharpe, W.F., 1971. Mean-absolute-deviation characteristic lines for securities and portfolios. Management Science 18 (2), B1–B13, October.
Sharpe, W.F., Tint, L.G., 1990. Liabilities—a new approach. Journal of Portfolio Management, 5–10, Winter.
Stulz, R., 1996. Rethinking risk management. Journal of Applied Corporate Finance 9, 8–24.
Stulz, R., 2001. Financial Engineering and Risk Management. Southwestern College Publishing.
Wind, Y., 1987. Financial services: increasing your marketing productivity and profitability. The Journal of Services Marketing 1, 5–18.
Zenios, S.A. (Ed.), 1993. Financial Optimization. Cambridge University Press, Cambridge, England.
Zenios, S.A., 1995. Asset/liability management under uncertainty for fixed-income securities. Annals of Operations Research 59, 77–98; Reprinted in: Ziemba, W.T., Mulvey, J.M. (Eds.), World Wide Asset and Liability Modeling. Cambridge University Press, Cambridge, England, 1998.
Zenios, S.A. 2006. Practical Financial Optimization: Decision Making for Financial Engineers. Basil Blackwell, Oxford, in press.
Zenios, S.A., Kang, P., 1993. Mean-absolute deviation portfolio optimization for mortgage backed securities. Annals of Operations Research 45, 433–450.
Ziemba, W.T., Mulvey, J.M., 1998. Worldwide Asset and Liability Modelling. Cambridge University Press, Cambridge, England.
Ziemba, W.T., Vickson, R.G. (Eds.), 1975. Stochastic Optimization Models in Finance. Academic Press, San Diego, CA; second ed. World Scientific, Singapore, 2006.

This page intentionally left blank

Chapter 2

TERM AND VOLATILITY STRUCTURES

ROGER J.-B. WETS

University of California, Davis, CA 95616, USA
e-mail: rjbwets@ucdavis.edu

STEPHEN W. BIANCHI

EpiSolutions Inc., El Cerrito, CA 94530, USA[1,2]
e-mail: swbianchi@earthlink.net

Contents

Abstract 26
Keywords 26
1. Term structure 27
  1.1. An example 27
  1.2. BootStrapping 30
    1.2.1. DFS-portfolio (deposits, futures and swaps) 31
    1.2.2. BootStrapping sub-portfolio 31
    1.2.3. BootStrapping results 32
    1.2.4. Generalized Bootstrap method 33
  1.3. Nelson–Siegel and Svensson's extension 34
  1.4. Maximum smoothness 36
  1.5. Forward-rates via geometric programming 37
  1.6. EpiCurves 38
    1.6.1. Zero-curves from spot rates 40
    1.6.2. Zero-curves from cash flow streams 42
    1.6.3. More examples 46
  1.7. A comparison for U.S. Treasury curves 51
2. Volatility structure 57
  2.1. Setting the stage 57
  2.2. Some tree-based valuation models 60
  2.3. The EpiVolatility model 61
    2.3.1. Guidelines 61

[1] Presently: Barra Inc., Berkeley, CA, USA.
[2] EpiSolutions Inc. now EpiRisk.

Handbook of Asset and Liability Management, Volume 1
Edited by S.A. Zenios and W.T. Ziemba
© 2006 Published by Elsevier B.V.
DOI: 10.1016/S1872-0978(06)01002-7

2.3.2. EpiSolutions Inc.'s approach 62
2.4. Implementation 62
2.5. Summary 65
Appendix A 65
References 67

Abstract

All valuations (discounted cash flow, instrument pricing, option pricing) and other financial calculations require an estimate of the evolution of the risk-free rates as implied by the term and volatility structures. This presumes that one has, if not perfect knowledge, at least very good estimates of these market term structures. In this chapter we review and compare the existing methodologies for deriving zero-curves (spot rates, forward rates and discount factors) and volatility estimates.

Keywords

spot and forward rate curve, discount factor curve, term structures, best fit, max-error criterion, error-sum criterion

JEL classification: G15

Term and volatility structures are the corner stone to practically all valuations of fixed income financial instruments and consequently affects, or should affect, significantly the trading and the management of financial holdings; they are also used to help shape monetary policy by Central Banks. Section 1 is devoted to a review and a comparison of the methods that have been suggested to construct the term structure associated with a given collection of fixed income financial instruments. Section 2 addresses the problem of determining the (associated) volatility structure that has been given only scant attention in the literature.

1. Term structure

1.1. An example

Let's begin with the following simple, but fundamental, issue: Find the zero-curves (= term structure) associated with a given portfolio. We use the term *zero-curve* in a generic sense to designate all or any one of the following financial curves: *spot rate, forward rate, discount factor and discount rate*; any zero curve completely determines the others. Indeed,

$$d_{t,m} = e^{-s_{t,m}m}, \quad s_{t,m} = m^{-1} \ln d_{t,m}, \quad f_{t,m} = s_{t,m} + m\dot{s}_{t,m},$$

where $s_{t,m}$ is the spot rate at time t for all bonds of maturity m, $d_{t,m}$ is the corresponding discount factor, and $f_{t,m}$ is the (instantaneous) forward rate a time t, again for bonds of maturity m; the upper dot on $\dot{s}_{t,m}$ stands for the time derivative.

Since the valuation of all fixed income financial instruments rest on starting with the 'right' zero-curves, it may come as a surprise that the zero-curves associated with a well-defined portfolio, derived using different methodologies, might vary significantly? This can best be illustrated by an example: On their web-site, TechHackers described a portfolio that includes Eurodollar Deposits, Eurodollar Futures and Swaps from June 10, 1997; see Tables A.1–A.3 in Appendix A. The seven pairs of *spot and forward rate* curves in Figure 1 are derived using

- three different implementations of BootStrapping,
- Svenson's extension of the Nelson–Siegel model,
- the updated Adams–Van Deventer maximum smoothness approach,
- Kortanek's derivation of the forward rates via geometric programming,

and finally, a novel approach based on Approximation Theory, implemented by EpiSolutions Inc.,

- relying on the construction of *EpiCurves*.

One might wonder if one should have any confidence in any one of these pairs! There are even noticeable differences between the spot rates curve. To understand the underlying reasons for these differences, one needs to examine the hypotheses under which these zero-curves were obtained. To do so, we begin with a description and an analysis of the BootStrapping methodology, to be followed by each one the alternatives listed

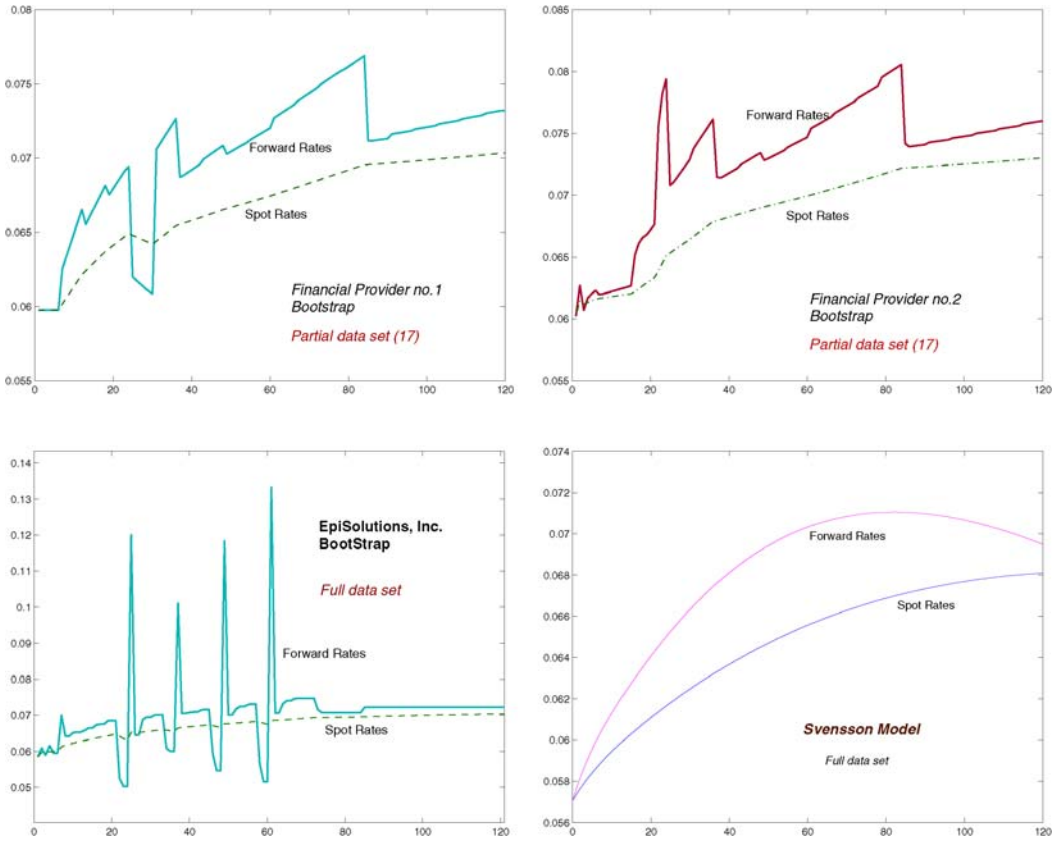


Fig. 1. Spot and Forward Rate curves: Seven different functionalities.

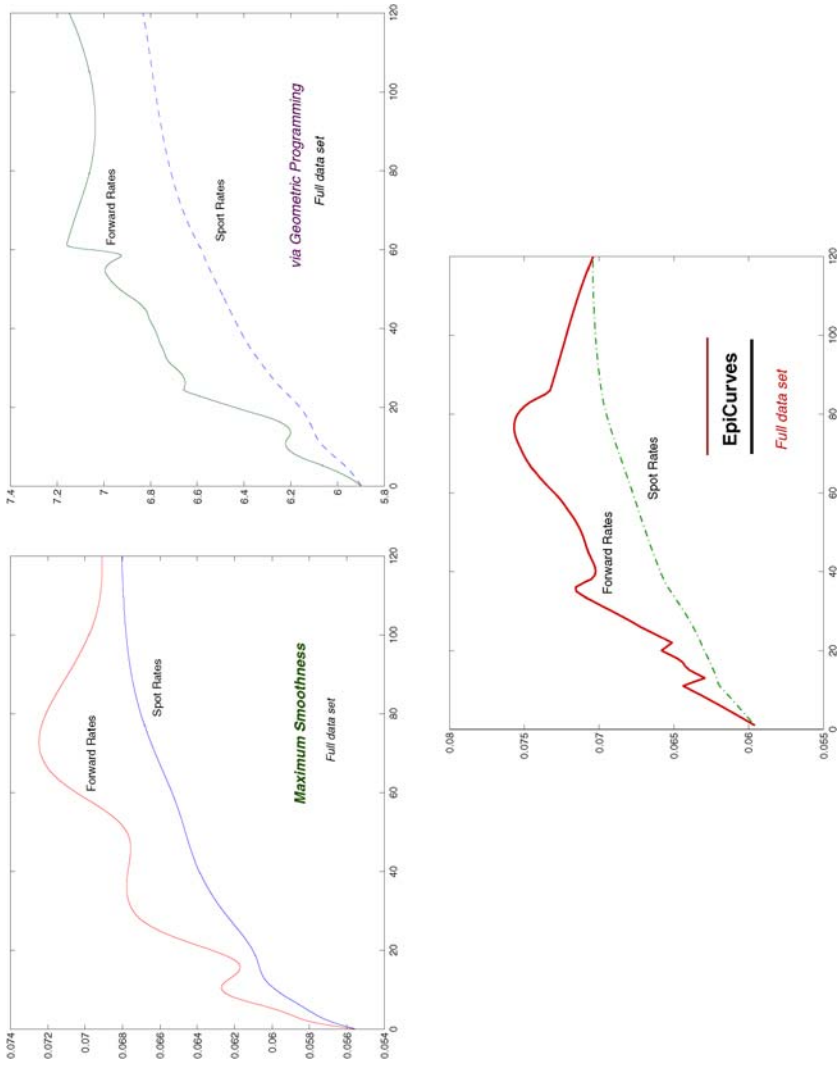


Fig. 1. *(Continued.)*

earlier. We shall then proceed with a discussion of the criteria that might be used in evaluating the 'quality' of the generated zero-curves and conclude with a comparison of the results based on these criteria.

1.2. BootStrapping

The valuation of fixed-income securities, and derivatives written on fixed-income securities, requires an estimation of the underlying risk-free term structure of interest rates. In principle, the term structure of interest rates is defined by a collection of zero-coupon bond prices (and their respective maturities), spanning the horizon over which a fixed-income security is to be valued. However, unless a zero-coupon bond exists for *every* maturity for which a discount factor is desired, some form of estimation will be required to produce a discount factor for any 'off-maturity' time. In practice, zero-coupon bond prices are available for a limited number of maturities (typically $\leqslant$1 year). If zero-coupon bonds for other maturities are available, a lack of liquidity may prevent the determination of an accurate or reliable price. As a result, the zero-curve is typically built from a combination of liquid securities, both zero-coupon and coupon-bearing, for which prices are readily available. This can include Treasury Bills, Deposits, Futures, Forward-Rate Agreements, Swaps, Treasury Notes and Treasury Bonds. The list need not be limited to these securities just noted, though most current vendor solutions *are* limited to these securities. Given a spanning set of securities, the zero-curve is then built using one of two forms of *BootStrapping*.

Under one BootStrapping method, the first step is to construct a larger set of spanning securities, by creating an 'artificial' security that matures on every date for which a cash flow is expected, and on which no security in the original set matures. For example, given a 5-year and a 6-year swap (paying semi-annually), a 5.5-year swap would be constructed, with a fixed-rate somewhere in between (based on some sort of interpolation) the fixed-rates for the 5-year and the 6-year swaps. Then, 'standard' BootStrapping may be applied to the expanded set of securities, giving discount factors for each maturity and cash flow date in the security set. This 'textbook' description of the BootStrapping method can be found in Hull (2000, §4.4) for example.

Another BootStrapping approach, is to make an assumption about how the instantaneous (or periodic) forward rates evolve between maturities in the security set. One assumption might be that forward rates stay constant between maturities, another might be that they increase or decrease in a linear fashion. Whatever the form of the forward rate evolution, *some assumption must be made*. Under this approach, instead of solving for a single discount factor for each successive security, a forward rate (or a parameter governing forward rate evolution) is obtained that will give the appropriate discount factor(s) between two maturity dates. For the example of the 5-year and 6-year swaps, given that the discount factors through the 5-year maturity have already been calculated, a forward rate is determined for the period between 5 and 6 years, that gives a 5.5-year discount factor and a 6-year discount factor that—when combined with the previous discount factors—will value the fixed side of the 6-year swap at par.

The results of either approach 'look' somewhat similar, a set of discount factors and corresponding dates spanning the horizon from today to the last maturity date in the security set. However, when using this set of discount factors as a basis for the valuation of other fixed-income securities, it will rarely be true that the cash flows of these securities will fall directly on the discount factor dates of the newly created zero-curve. In this case, a discount factor or a zero-coupon rate must be interpolated from the spot-curve. Typically available interpolation methodologies for this include linear, log-linear, exponential, cubic-spline, or any of a number of variations on fitting the zero-curve with a polynomial.

As long as two securities do not share the same maturity date, any combination of securities may theoretically be used in constructing a zero-curve, even with the BootStrapping methodology. A limitation of the currently available technology, is that the user *must* typically 'switch' from one security type to another during the BootStrapping process. For example, given a set of deposits, futures, and swaps, the currently available methods will not allow for inclusion of a deposit and a futures contract, where the underlying deposit maturity date of the futures contract is *prior* to the maturity date of the deposit (similarly, no 'overlap' is allowed between futures contracts and swaps). At first one might not consider this as a serious limitation, as users may very well wish to describe different 'sections' of the zero-curve using certain types of securities. Almost all web-based, commercially available zero-curve construction technologies rely on a form of *BootStrapping*, in combination with a variety of interpolation methods. Furthermore, they are limited to using certain types of securities, and are also limited in the ways in which these securities may be combined.

1.2.1. DFS-portfolio (deposits, futures and swaps)

We now return to the example found on the TechHackers web-site. A detailed description of the portfolio is given in Tables A.1–A.3 in Appendix A, but at this point at brief description will suffice. There are 36 instruments in the full data set, broken down as follows: 6 Eurodollar Deposits with term-to-maturities ranging from overnight to 12 months, 24 Eurodollar Futures with 90-day deposit maturities ranging from 3 months to 6 years, and 6 Swaps with term-to-maturities ranging from 2 to 10 years. The yields of the deposits vary from 5.55 to 6.27% with the higher yields corresponding to those with larger maturity. Similarly, the 24 futures have yields that vary from 5.89 to 7.27%. The yields for the swaps vary from 6.20 to 6.86%, again with the yields increasing as maturities get larger.

1.2.2. BootStrapping sub-portfolio

In this example, there is overlap between the maturity dates for the EuroDollars Deposits (1 day, 1 month, 2 m., 3 m., 6 m., 12 m.) and the EuroDollars Futures (3 months, 6+ m., 9 m., 12 m., 15 m., ...). Thus *not all instruments can be included* in a 'BootStrapping sub-portfolio'. One possible choice, the default setting for Financial Provider No. 1, is

to switch from Eurodollar Deposits to Eurodollar Futures at the earliest possible time, i.e., include the first available Futures contract. Similarly, switching from Eurodollar Futures to Swaps occurs at soon as possible, i.e., at a time corresponding to the Swap with the lowest maturity. With this selection criteria, the data set used to generate the 'BootStrapping sub-portfolio' is a subset of the full DFS-portfolio data set that includes: the first five Eurodollar Deposits, the 2nd through the 7th Eurodollar Futures contracts and all 6 Swaps—for a total of 17 instruments.

The user may choose any time they wish for making these switches, the point is that they *must be made somewhere* along the line. Clearly, Financial Provider No. 2 relied on a different selection criterion to constitute its BootStrapping sub-portfolio.

1.2.3. BootStrapping results

The first two pairs of spot and forward rate curves in Figure 1 are those derived from functionalities made available by two Financial Providers. Although all the spot rate curves appear to be relatively similar (except for the time span from month 20 to month 25), the forward rate curves are quite dissimilar. Of course, this can be traced back to the different BootStrapping implementations that rely on the selection of different sub-portfolios, as indicated earlier.

The implementation of the BootStrapping technique at EpiSolutions Inc.—based on the simple precept that the instantaneous forward rates are constant between (adjacent) maturity dates—for the same BootStrapping sub-portfolio as Financial Provide No. 1 yields the pair of zero-curves in Figure 2; these spot and forward rate curves are similar to those obtained by Financial Provider No. 1. The same approach, but with the *full*

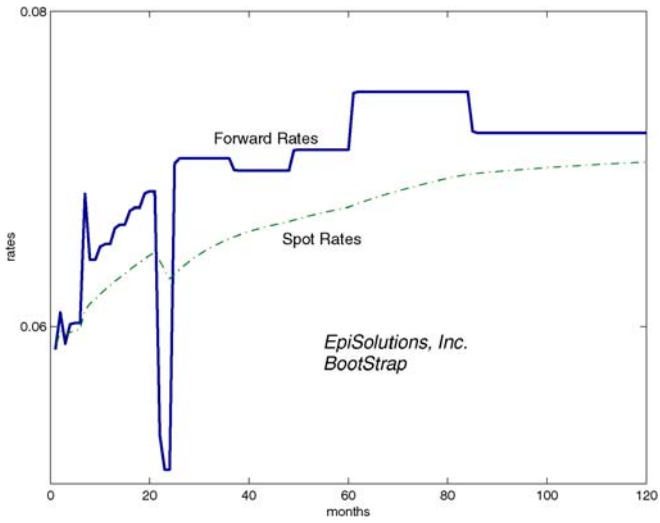


Fig. 2. ESI-BootStrapping for the BootStrapping sub-portfolio.

data set, i.e., when there is no mutilation of the given DFS-portfolio, yields the pair of zero-curves in the first graph of line 2 of Figure 1.

1.2.4. Generalized Bootstrap method

Deaves and Parlar (1999) suggested an approach that overcomes the need to mutilate a portfolio to be able to apply the 'textbook'-version of BootStrapping described above. Their approach, unusual and interesting, consists in calculating the spot-rates r_t associated with each one of the time-dates when there is a cash flow (of any type). In our example, this corresponds to coupon-, settling-, delivery- and maturity-dates; in fact, there are 73 such dates. However, there are only 36 instruments that would generate the following 36 pricing-out equations

$$0 = \sum_{t\in\mathcal{T}} \frac{1}{2} p_{kt} e^{-r_t \cdot t}, \quad \forall k = 1, \ldots, 36,$$

where the instruments are indexed by k. $\mathcal{T}$ is an ordered set of all the time-dates at which cash flow occurs and p_{kt} is the cash flow (positive or negative) for instrument k at time t. It's also assumed that there no repeated maturity-dates for the instruments in the (given) portfolio. This non-linear system is solvable, but has 37 more variables than equations and, consequently, will have multiple solutions. In order to generate a system with (hopefully) a unique solution, Deaves and Parlar add 37 linear equations that relate the rates at the 37 time-dates that are not maturity-dates, say $r_{37}, \ldots, r_{73}$, to the rates at maturity-dates: say $r_1, \ldots, r_{36}$. To do so, they rely on a Maple (symbolic) cubic spline fit functionality; this cubic spline, $cs(\cdot)$ consists of 36 different cubics: one for each of the time-intervals between the different, increasing maturity dates. For each one of these time intervals,

$$cs(t) = \sum_{j=1}^{3} 6c_{0j} r_j + \left(\sum_{j=1}^{3} 6c_{1j} r_j\right) t + \left(\sum_{j=1}^{3} 6c_{2j} r_j\right) t^2 + \left(\sum_{j=1}^{3} 6c_{3j} r_j\right) t^3,$$

where the 144 coefficients $\{c_{lj},\ l = 0, \ldots, 3,\ j = 1, \ldots, 36\}$ are those calculated by the (symbolic) cubic spline fit. From this, one can obtain a linear equation for each one of the 37 non-maturity-dates rates. For our example, the fourth cash flow non-maturity-date is 0.5147 years after June 10, 1997 ($t = 0$) and the corresponding linear equation is

$$\begin{aligned} r_{40} = {} & .0000356 r_1 - .0002056 r_2 + .0013674 r_3 - .0124579 r_4 + .015465 r_5 \\ & - .322488 r_6 + 1.3053364 r_7 + .0162911 r_8 - .0133038 r_9 + .01005185 r_{10} \\ & - .0001131 r_{11} + .0000272 r_{12} - .768069 e^{-5} r_{13} + .783936 e^{-5} r_{14} \\ & - .606228 e^{-5} r_{15} + .556858 e^{-7} r_{16} - .1349305 e^{-7} r_{17} + .379412 e^{-8} r_{18} \\ & - .4295145 e^{-8} r_{19} + .3402209 e^{-8} r_{20} - .2493217 e^{-10} r_{21} + .61491 e^{-11} r_{22} \\ & - .174375 e^{-11} r_{23} + .2176 e^{-11} r_{24} - .1767913 e^{-11} r_{25} + .998711 e^{-14} r_{26} \end{aligned}$$

$$
\begin{aligned}
&- .2495984e^{-14}r_{27} + .708697e^{-15}r_{28} - .1002684e^{-14}r_{29} \\
&+ .835819e^{-15}r_{30} - .352076e^{-17}r_{31} + .88496e^{-18}r_{32} - .2281451e^{-18}r_{33} \\
&+ .420647e^{-19}r_{34} - .1130008e^{-20}r_{35} + .414996e^{-22}r_{36}.
\end{aligned}
$$

The 37 equations allow us to replace the 37 non-maturity-date rates in the 36 basic (non-linear) equations and reduce the system to one involving only 36 equations and 36 unknowns! Such a system can be solved by relying on either Newton's method, one of its variants (Quasi-Newton methods), or any other appropriate method; the only drawback is that the Jacobian of the resulting system would be dense and it would not be possible to take advantage of sparsity; Deaves and Parlar rely on Maple's `fsolve` function.

The solution of this system yields the 73 spot-rates at all 'cash-flow' dates. Some further interpolation(s) would be required to obtain the yield curve at all times 't'. Although, Deaves and Parlar do not suggest any specific interpolation strategy, presumably, one would rely on the cubic spline cs, after substituting the values for $r_1, \ldots, r_{36}$, to obtain the spot rates at any time t; forward rates and discount factors would be derived, in turn, from this spot rate curve.

We tried to implement this approach following the steps suggested by Deaves and Parlar, i.e., by relying on the functionalities provided by Maple. Such an implementation is labor intensive, and after a couple of days, we realized that for this relatively simple 36 instruments example, from modeling to the point where you can input the 36-by-36 (non-linear) system might very well take a week's work. That is expensive and the delay would be much too long for most potential applications. Moreover, even for this example, the time required to derive the (symbolic) cubic spline was non-negligible. The idea of dealing in this manner with a portfolio with 100+ instruments and 2000 to 3000 cash-flow dates, is to say the least, intimidating; there is also the requirement that no instruments can have the same maturity-dates. *All* these obstacles can be overcome but to do so one needs to get involved in a *major* implementation effort and rely on much more sophisticated procedures to obtain the cubic spline (cs) and to solve (dense) non-linear systems than the `spline` and `fsolve` Maple-functionalities. Because we could not handle portfolios of reasonable (practical) size, it has not been possible to include this intriguing approach in our analysis.

1.3. Nelson–Siegel and Svensson's extension

Under all the limitations mentioned in the previous section, BootStrapping strives to accomplish the (presumably) desirable goal of pricing the securities used in the zero-curve estimation exactly. In most cases, however, it does not produce believable forward rates and even the spot rate curve fails to be smooth, the norm for such curves. There can be many technical reasons for this, that are not inherent to the BootStrapping method itself, such as liquidity, tax effects, and/or missing data points. Regardless, practitioners would like a method of estimating zero-curves that is robust and reasonable, across many different markets and market conditions. One criterion that is especially desirable

for those using zero-curves for strategic planning purposes, is *smoothness*, particularly with regard to the evolution of the implied forward-rate curve. There are many ways to estimate zero-curves whose associated forward-rate curves are smooth, one of which is to posit a functional form that is smooth by definition, and try to find best-fit parameters for this function. In 1987, Nelson and Siegel (1987) did just this by proposing the following formula for the evolution of the instantaneous forward rate (at time t)

$$f_{t,m} = \beta_{t,0} + \beta_{t,1} e^{-m/\tau_{t,1}} + \beta_{t,2}(m/\tau_{t,1}) e^{-m/\tau_{t,1}}$$

where m is the time-to-maturity and $\beta_{t,0}, \beta_{t,1}, \beta_{t,2}, \tau_{t,1}$ are the parameters to be estimated to fit as well as possible the available data. Integrating the forward rate curve —and dropping the index t—yields the spot rate

$$s_m = \beta_0 + (\tau_1/m)(\beta_1 + \beta_2)(1 - e^{-m/\tau_1}) - \beta_2 e^{-m/\tau_1}.$$

Typically, this expression for the spot rates will generate a curve with one 'hump', to allow for a second 'hump' Svensson (1995) proposed an extension to this formula that generally increased the flexibility for fitting a given set of securities and their market prices

$$f_m = \beta_0 + \beta_1 e^{-m/\tau_1} + \beta_2(m/\tau_1) e^{-m/\tau_1} + \beta_3(m/\tau_2) e^{-m/\tau_2}$$

where $\beta_0, \beta_1, \beta_2, \tau_1, \beta_3, \tau_2$ are the parameters to be estimated. The corresponding spot rate curve is

$$s_m = \beta_0 + \beta_1 \frac{-\tau_1}{m}(1 - e^{-m/\tau_1}) + \beta_2\left(\frac{-\tau_1}{m}(1 - e^{-m/\tau_1}) - e^{-m/\tau_1}\right) + \beta_3\left(\frac{-\tau_2}{m}(1 - e^{-m/\tau_2}) - e^{-m/\tau_2}\right).$$

One or the other of these two formulas is in use at a large number of central banks (BIS, 1999), where the resulting zero-curves are used to help shape monetary policy; clearly, a strategic planning activity.

Both the Nelson–Siegel and the Svensson models yield smooth zero-curves. That, and the fact that one can explicitly calculate spot and forward rates at any time in terms on a relatively simple formula are certainly desirable attributes. But, these are also the basis for the shortcomings of this approach. The fact that the rate curves come with a precise analytic expression that depends on either four or six parameters might very well result in a 'fit' that does not take full account of the market structure, i.e., it will be difficult to match market prices with a high level of precision: there is an inherent lack of flexibility.

Moreover, because the fitting these parameters leads to an highly non-convex optimization problem, one essentially has to resort to global optimization techniques that are somewhat unreliable unless the search is exhaustive. Equivalently, one could look for a 'good' critical point by solving repeatedly a collection of nonlinear systems by relying on some heuristics to trim down the number of systems that need to be considered.

This is the approach suggested by Svensson and implemented in various institutions (BIS, 1999); a description of our implementation can be found in Section 1.7. Thus, the downside of this numerical approach is, on one end, the uncertainty about having considered a sufficient number of (discrete) possibilities and, on the other end, the need for the user to 'intervene' appropriately usually resulting in a time consuming operation.

1.4. Maximum smoothness

The definition of smooth for a curve might almost be subjective, but it's certainly application dependent. In the world of zero-curves, one is certainly not going to be satisfied with a 'smooth' spot rate curve, or a 'smooth' discount factor curve, when an associated zero-curve, for example the forward rate curve, is seesawing; that this can actually occur was already pointed out in Shea (1985). In fact, it suffices return to Figure 1 and look at the Forward Rate curves generated by BootStrapping, even those derived from the 17-instruments sub-portfolio.

Adams and Van Deventer (1994) rely on a criterion used in engineering applications, cf. Hildebrand (1987), in their derivations of spot and forward rate curves with 'maximum smoothness'. They propose finding a forward rate curve, fw, such that for each instrument in the portfolio:

$$P_i = \exp\left(-\int_0^{t_i} fw(s)\,\mathrm{d}s\right), \quad i = 1, \ldots, I,$$

where P_i is the (today's) price and t_i the maturity date of instrument i, and

$$\int_0^T \left[fw''(s)\right]^2 \mathrm{d}s \text{ is minimized,}$$

$[0, T]$ is the time span in which we are interested; usually T is the largest maturity of the instruments in the portfolio. It is shown that the solution is a 4th order spline, certainly smooth, whose coefficients can be easily computed; 'maximum' smoothness is achieved in terms of a criterion attributed to Vasicek.

The Achilles' heel of this approach, at least as laid out in Adams and Van Deventer (1994), is that the only instruments that can be included in the 'maximum smoothness' portfolio are zero-coupon bonds, and zero-coupon bonds with maturities exceeding one year are extremely rare. In order to obtain zero-curves that span more than a few months, one possibility is to fabricate artificial (long term) zero-coupon bonds that have similar financial characteristics to those instruments found in the portfolio; this requires interpolations of some type. Moreover, prices P_i are present day prices and so no future contracts can be included in the 'maximum smoothness' portfolio. Presumably, this can also be skirted by some adjustments. In the final analysis, like for 'standard' BootStrapping, we have to create a sub-portfolio and then enrich it by artificial instruments in order to be able to apply the suggested method.

From our previous examples and analysis, it's clear that if one is going to derive zero-curves by taking into account more than just a few well chosen instruments, and

one is going to aim at an acceptable level of smoothness, there is going to be a 'price' to pay for this. In Function Theory, the smoothness of a curve is identified in terms of the number of times it's continuously differentiable. A curve $z : [0, T] \to \mathbb{R}$ is said to be of class $\mathcal{C}^q$ if its qth derivative is continuous. So, if z is of class $\mathcal{C}^2$ it means that it can be differentiated twice and the second derivative is continuous. If it's of class $\mathcal{C}^0$ then z is just continuous, and if it's of class $\mathcal{C}^\infty$ then all derivatives, of any order, exist and are continuous. It is evident that $\mathcal{C}^0 \supset \mathcal{C}^1 \supset \cdots \supset \mathcal{C}^\infty$. One might wish the zero-curves to be of class $\mathcal{C}^\infty$, but it is clear that this is a much smaller family of curves than those that are just continuous, or just continuously differentiable: $\mathcal{C}^1$. Consequently, by insisting that our zero-curve be of class $\mathcal{C}^\infty$, we might very well have excluded those curves that have the 'accuracy' properties we are looking for. Hence, we usually must be content with smooth curves that are less than *infinitely* smooth.

1.5. Forward-rates via geometric programming

A completely novel and tantalizing approach was developed by Kortanek (2003), in collaboration with Medvedev (Kortanek and Medvedev, 1999, 2001) and Yunianto (Kortanek and Yunianto, 2004), that's focused on obtaining the forward rate curve. Their motivation, certainly in part, came from the shortcomings of the Nelson and Siegel, and Svensson models. In Kortanek and Medvedev (1999, 2001), a forward rate model was developed allowing for non-stochastic perturbations

$$\frac{\mathrm{d}fw(t)}{\mathrm{d}t} = \alpha + \beta fw(t) + v(t), \quad fw(0) = r_0,\ t \in [0, T],$$

where r_0, α, β and the perturbation function $v : [0, T] \to \mathbb{R}$ are the parameters that need to be estimated; the interval $[0, T]$ could be unbounded. The function v is assumed to be piecewise constant, i.e.,

$$v(t) = v_i \quad \text{if } t \in (t_{i-1}, t_i]$$

where the subintervals $(t_{i-1}, t_i]$ partition $(0, T]$. For $t \in (t_{i-1}, t_i]$, one has (Kortanek and Medvedev, 2001, (9.12)),

$$fw(t) = r_0 e^{\beta t} + \frac{\alpha}{\beta}(e^{\beta t} - 1) + \frac{e^{\beta t}}{\beta} \sum_{j=1}^{i-1} (e^{\beta t_j - 1} - e^{\beta t_j}) v_j + \frac{e^{\beta(t - t_{j-1})} - 1}{\beta} v_j.$$

If the interval $[0, T]$ has been partitioned in I intervals, fitting the forward rate curve will require estimating $I + 3$ parameters. In setting up this estimation problem, one can introduce a number of 'natural' bounds on all or some of the variables; for example, requiring $\beta \in [\beta_l, \beta_u] \subset \mathbb{R}_-$ would result in a mean reversion property for the spot rates.

Since this formula provides us with the forward rates for any time t, the price of any bond generating a given cash flow, coupons and notional with maturity less than or equal to T, can be expressed in terms of these parameters, cf. Kortanek and Yunianto (2004, §3.2). The expression gets a bit involved because one has to take into account

coupons that get paid a times that fall between the end points of the intervals $(t_{i-1}, t_i]$, but this is no more than keeping the 'bookkeeping' straight. Anyway, one ends up with an expression that is a *posynomial* of the following type:

$$P(t_k) = e^{a_k r_0} e^{b_k \alpha} \left(\prod_{j=1}^{J_k} e^{a_j v_j} \right) e^{b_{J_k} v_{J_k}} + \left(q_k \sum_{l=1}^{n_k} e^{a_{kl} r_0} e^{b_{kl} \alpha} \right) \prod_{j=1}^{J_k - 1} e^{c_j v_j} e^{a_{kl} v_j} e^{b_{J_k l} v_{J_k}}$$

for a bond of notional 1 and maturity t_k that falls in the interval $(t_{J_k-1}, t_{J_k}]$, n_k is the number of coupons before maturity, and the 'coefficients' $a_j, b_j, \ldots$ are themselves functions of α and β; precise definitions can be found in Kortanek (2003) and Kortanek and Yunianto (2004, §3 & §4).

The strategy is now to find a best fit between the estimated prices $P(t_k)$ and the observed prices $\widehat{P}(t_k)$ that respects the *no-arbitrage* conditions. Taking advantage of some elegant simplifications, Kortanek show that the problem to be solved is a *geometric program*, i.e., an optimization problem whose objective and constraints are posynomials. This would fall in the extended family of convex optimization problems except for the presence of so-called 'reversed constraints' that are part of the formulation if the collection of securities being considered includes coupon paying bonds. Formulating a practical problem as a geometric program is called *GP modeling*. "GP modeling is not just a matter of using a software package or trying out some algorithm; it involves some knowledge, as well as creativity, to be done effectively (Boyd et al., 2004)." This is carried out effectively by Kortanek et al., cf. Kortanek and Yunianto (2004).

We have not implemented ourselves this procedure but have relied on the results provided by K. Kortanek, cf. Figure 1. On a number of test problems, this approach came the closest to the quality of the zero-curves generated by the method described in the next section, but our lack of an independent implementation means that we could not include it in the comparisons of Section 1.7. Further results, variants and extensions of this method can be found in Kortanek and Yunianto (2004).

1.6. EpiCurves

EpiCurves, introduced in Wets, Bianchi and Yang (2001), come from a large, but specific sub-family of curves that are of class $\mathcal{C}^q$ for some $q = 1, 2, \ldots$. EpiCurves could be viewed as 'constrained' splines, however their derivation does not follow the standard spline-fitting techniques. To simplify the presentation, suppose we are interested in finding a $\mathcal{C}^2$-curve z on $[0, T]$ given by

$$z(t) = z_0 + v_0 t + \frac{1}{2} a_0 t^2 + \int_0^t \int_0^\tau \int_0^s x(r)\, \mathrm{d}r\, \mathrm{d}s\, \mathrm{d}\tau, \quad t \in [0, T],$$

where

- $x : (0, T) \to \mathbb{R}$ is an arbitrary piecewise continuous function that corresponds to the 3rd derivative of z;
- a_0, v_0, z_0 are constants that can be viewed as integration constants.

Once the function x (3rd derivative) and the constants a_0, v_0, z_0 have been chosen, the function z is completely determined.

Let's now go one step further. Instead of allowing for any choice for x, let's restrict the choice of x to piecewise constant functions of the following type: split $[0, T]$ into N sub-intervals of length T/N and let the function x be constant on each one of these intervals, with

$$x(t) = x_k \quad \text{when } t \in (t_{k-1}, t_k],\ k = 1, \ldots, N,$$

where $t_0, t_1, \ldots, t_L$ are the end points of the N sub-intervals. The corresponding curve z on $[0, T]$ is completely determined by the choice of

$$a_0, v_0, z_0 \quad \text{and} \quad x_1, x_2, \ldots, x_N,$$

i.e., by the choice of a finite number of parameters, exactly $N + 3$.

For $k = 1, \ldots, N$, $t \in (t_{k-1}, t_k]$ with $\delta = T/N$ and $\tau = t - t_{k-1}$,

$$\begin{aligned}
z(t) &= z_0 + v_0 t + \frac{1}{2}a_0 t^2 + \frac{\tau^3}{6}x_k \\
&\quad + \sum_{j=1}^{k-1} x_j \left[\frac{\delta^3}{2}\left(\frac{1}{3} + (k-j-1)(k-j)\right) + \delta^2\tau(k-j-0.5) + \frac{\delta\tau^2}{2}\right], \\
z'(t) &= v_0 + a_0 t + \sum_{j=1}^{k-1} x_j\left(\delta^2(k-j-0.5) + \delta\tau\right) + \frac{\tau^2}{2}x_k, \\
z''(t) &= a_0 + \delta\sum_{j=1}^{k-1} x_j + \tau x_k, \\
z'''(t) &= x_k.
\end{aligned}$$

By restricting the choice of x to piecewise constant functions, the resulting z-curves are restricted to those curves in $\mathcal{C}^2$ that have (continuous) piecewise linear second derivatives. Designate this family of curves by $\mathcal{C}^{2,pl}$ where *pl* stands for piecewise linear; whenever appropriate we use the more complete designation $\mathcal{C}^{2,pl}([0, T], N)$ with $[0, T]$ the range on which these curves are defined and N the number of pieces, but usually the context will make it evident on which interval these curves are defined. Clearly, not all $\mathcal{C}^2$-curves are of this type. However, Approximation Theory for functions, tells us that any $\mathcal{C}^2$-curve can be approximated *arbitrarily closely* by one whose second derivative is a continuous piecewise linear function, i.e., a curve in $\mathcal{C}^{2,pl}([0, T], N)$, by letting $N \to \infty$. This provides us with the justification one needs to restrict the search for 'serious' zero-curves to those in this particular sub-family of $\mathcal{C}^2$-curves.

Later on, the implementation will impose further restrictions on the choice of the coefficients, not to guarantee 'smoothness' in itself since every curve in $\mathcal{C}^{2,pl}$ is clearly (mathematically) smooth, but to generate zero-curves that would be called 'smooth' by a practitioner.

In summary, the building of *EpiCurves* starts by selecting the level of smoothness desired ($z \in \mathcal{C}^q$), and then a zero-curve is built whose qth derivative is a continuous piecewise linear function. This requires fixing a finite number of parameters; actually $N+q+1$ parameters. If the resulting curve does not meet certain accuracy criteria, the step size (T/N) is decreased by letting $N \to \infty$.

1.6.1. Zero-curves from spot rates

To set the stage for finding the zero-curves associated with a collection of instruments generating cash flow streams, let's consider an *EpiCurves* approach to fitting spot rates to obtain a spot rates (yield) curve. The data come in a pair of arrays,

$$s = (s_1, s_2, \ldots, s_L), \quad m = (m_1, m_2, \ldots, m_L),$$

that give us the spot rates for a collection of instruments of different maturities, for example, Treasury Notes. The task is to find a spot rate curve that 'fits' these data points. That is easy enough. Assuming that $m_1 < m_2 < \cdots < m_L$, one could simply derive a spot rate curve by linear interpolation between adjacent pairs. That's actually a perfect fit. Generally, this is not a 'smooth' curve. This usually generates a forward rate curves that can be quite jagged. So this 'simple' solution almost never produces zero-curves that practitioners would consider acceptable. Of course, one can use another interpolation method, such as via quadratic or cubic splines, that generates significantly better results. Another possibility is to set-up an artificial portfolio with coupon-bonds whose yields would match the given spot rates. The problem is then reduced to one of finding the zero-curves associated with the cash flow stream of this (artificial) portfolio. This is dealt with in the next subsection. But, this latter approach, found in the packages of some financial technology providers, circles around the problem, at least one too many times, before dealing with it.

The use of the *EpiCurves* technology provides an elegant solution that generates smooth zero-curves. The strategy is to find a spot rate curve of the type described in the previous subsection, say again a $\mathcal{C}^{2,pl}$-curve that will match the given spot rates. One must accept the possibility that we won't be able to find, for a fixed N, a $\mathcal{C}^{2,pl}([0,T],N)$-curve that fits perfectly the given data. So, the problem becomes one of finding the 'best' possible fit. Best possible can be defined in a variety of ways but it always comes down to minimizing the 'error', i.e., the distance between the *EpiCurves* and the given spot rates. Mathematically, the problem can is

$$\text{find } z \in \mathcal{C}^{2,pl}\big([0,T],N\big) \text{ so that } \big\|s - z(m_1 : m_L)\big\|_p \text{ is minimized,}$$

where $z(m_1 : m_L) = (z(m_1), z(m_2), \ldots, z(m_L))$ and $\|a\|_p$ is the ℓ^p-norm of the vector a. With $p = 1$, one would be minimizing the sum of the (absolute) errors, with $p = 2$ one minimize the sum of the squares of the errors, and with $p = \infty$, it would be the maximum (absolute) error that would be minimized. An implementation by EpiSolutions Inc., has $p = 1$ and thus minimizes the sum of the errors, since

$$\|s - z(m_1 : m_L)\|_1 = \sum_{l=1}^{L} |s_l - z(m_l)|.$$

The resulting optimization problem can then be reduced to a linear programming problem, since, as explained in the previous subsection, the functions z in $\mathcal{C}^{2,pl}$ are completely determined by a finite number of parameters.

To illustrate the results, we apply both linear interpolation and the *EpiCurves* technology to obtain a spot rate curve that fits the spot rates (for T-bills and Treasury notes) of October 1982:

$$m = (3, 6, 12, 24, 36, 60, 84, 120, 240, 360),$$

$$s = (7.97, 8.63, 9.32, 10.19, 10.62, 10.8, 10.88, 10.91, 10.97, 11.17);$$

the time unit is 1 month. The spot and forward rate curves can be found in Figure 3. It's barely possible to see the difference between the spot rate curves, but the difference between the forward rate curve is more than noticeable. The difference, of course, can be traced back to the intrinsic smoothness of the spot rate curve when it's generated as an *EpiCurve*.

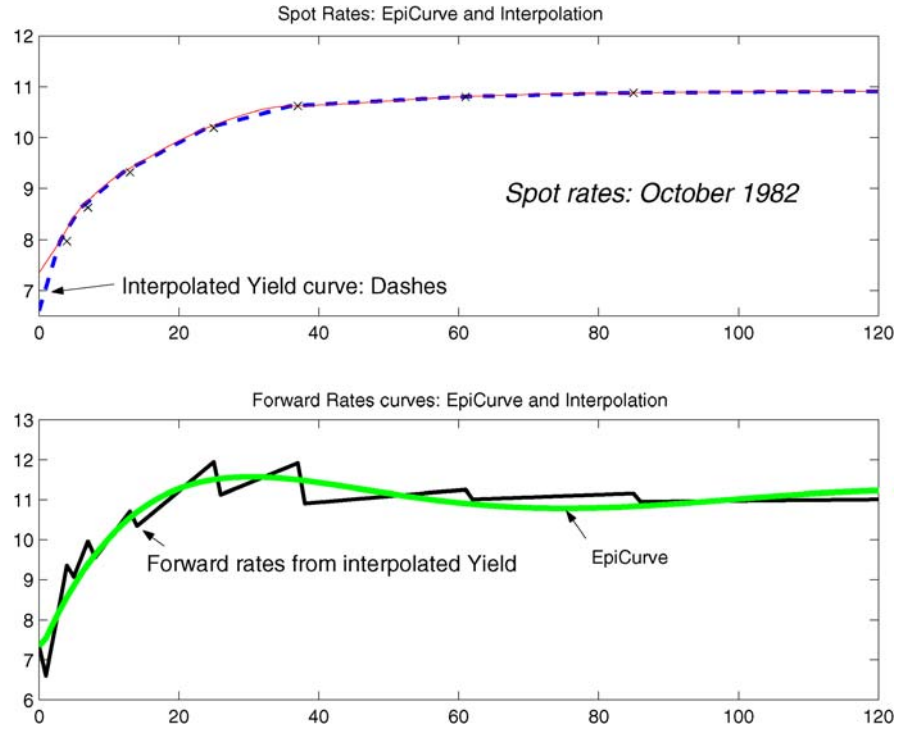


Fig. 3. Linear interpolation vs. EpiCurves.

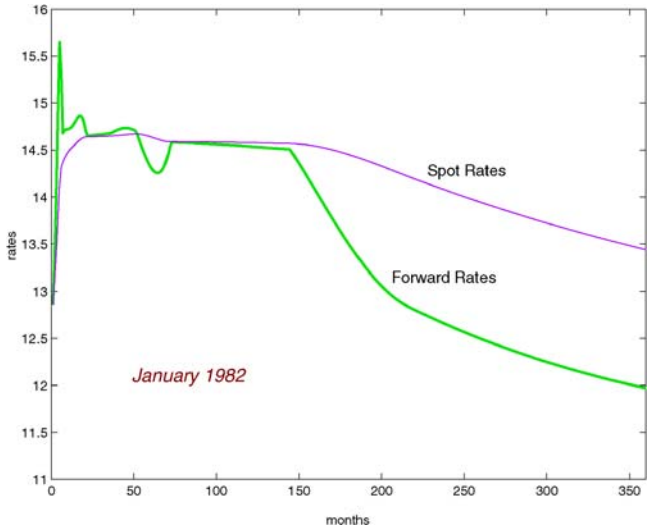


Fig. 4. Spot and Forward Rate curves from spot rates.

Let also consider the case spot rates for January 1982, the maturities-array m is the same, but now

$$s = (12.92, 13.90, 14.32, 14.57, 14.64, 14.65, 14.67, 14.59, 14.57, 14.22).$$

Running *EpiCurves* yields the result in Figure 4. The forward rate curve is rather unsettled up to the end of year 1, it actually reflects almost perfectly the 'unsettled' market situation at that time (January–February 1982).

1.6.2. Zero-curves from cash flow streams

We briefly review our goals and guiding principles. Given the increased complexity of the instruments being traded, it certainly is no longer sufficient to be able to build zero-curves based on just zero-coupon bonds; in fact, the LIBOR (swaps) based zero-curves seem to occupy, at present, the prominent place. The ultimate objective should be to build the spot rate curve associated with *any* collection of instruments, for example, AAA- or AA-rated corporate bonds, any mixture of swaps, futures and bonds, etc. Notwithstanding a relatively large literature devoted to zero-curves, cf. Buono, Gregoru-Allen and Yaari (1992), there has never been any serious attempt at dealing with the building of zero-curves at this more comprehensive level.

Of course, given an arbitrary collection of instruments, each one generating its own cash flow stream, it might be possible (assuming that maturities occur at different dates) to generate, via BootStrapping for example, any one of the zero-curves. However, as every practitioner knows all too well and as was reviewed in Section 1.2, some of the

resulting curves will be, to say the least, unwieldy, and have every characteristic except 'believable'. The insistence on 'smoothness', cf. Vasicek and Fong (1977), Shea (1985), Adams and Van Deventer (1994), is motivated by the strongly held belief, that's also supported by historical data, that zero-curves do not come with kinks, and spikes i.e., extremely abrupt changes in the rates.

Keeping this in mind, the problem of generating zero-curves could be roughly formulated as follows: Given a collection of instruments, each one generating a given cash flow stream, find smooth zero-curves so that for each instrument (in the collection), the *net present value* (*NPV*) of the associated cash flow matches its present price.

Although this formulation allows us to include zero-coupon bonds, coupon bonds, swaps, etc., in our collection of instruments, it does not allow for futures, future swaps, etc. To do so, we reformulate the problem in the following more general terms: With each instrument i in our collection, we associate a

$$\text{Time array:} \quad (t_{i1}, t_{i2}, \ldots, t_{i,L_i}),$$

the dates, or time span, at which cash payments will take place, and a

$$\text{Payments array:} \quad (p_{i1}, p_{i2}, \ldots, p_{i,L_i})$$

with cash flow p_{il} received at time t_{il}, L_i is the maturity date. One then interprets $p_{il} > 0$ as cash received and $p_{il} < 0$ as cash disbursed. For example, in the case of a coupon bond, bought today for \$100, with semi-annual \$3 coupons and a two year-maturity, one would have:

$$\text{Time array:} \quad (0, 6, 12, 18, 24),$$

assuming the time unit is '1 month', and

$$\text{Payments array:} \quad (-100, 3, 3, 3, 103).$$

This allows us to include almost any conceivable instrument in our collection, as long as it comes with an explicit cash flow stream. For example, in the case of the following T-bill forward: A bank will deliver in 3 months from now, a 6-month Treasury bill of face value \$100 with a 10% annual forward rate for that 6 months period. The value of such a contract would be \$95.24 that would have to be paid in 3 months. This contract would then come with the following arrays:

$$\text{Time array:} \quad (3, 9), \qquad \text{Payments array:} \quad (-95.24, 100).$$

In this frame of reference, the zero-curve problem could be formulated in the following terms: Given a (finite) collection of instruments that generate cash flow streams, find a *discount factor curve* such that

- the net present value (NPV) of each individual instrument (contract) turns out to be 0 when all cash payments received and all disbursements are accounted for;
- all associated zero-curves (forward, spot, discount rate) are 'smooth'.

When formulated at this level of generality, the zero-curve problem is usually not feasible. In fact, it's not difficult to fabricate an 'infeasible' problem. Simply, let the collection consist of two one-coupon bonds that have the same nominal value, the same maturity and the same price (today). Both coupons are to be collected at maturity but have different face value. Clearly, there is no discount factor curve so that the net present value (NPV) of both of these cash flows turns out to be 0! Of course, this is an unrealistic example, the financial markets would not have assigned the same price to these two instruments; arbitrage would be a distinct possibility in such a situation. But since we allow for *any* collection of instruments, there is the distinct possibility that there are practical instances when one cannot find a 'smooth' discount factor curve so that the NPV of *all* cash flow streams factors out to 0. So, given that we want to be able to deal with any eclectic collection of instruments, as well as the 'standard' ones, instead of asking for the NPV of all cash flow streams to be 0, we are going to ask that they be as close to 0 as possible.

Smoothness is going to be achieved by restricting the choice of the discount factor curve to $\mathcal{C}^{q,pl}$, i.e., curves whose qth derivative is continuous and piecewise linear as introduced in this section. To render our presentation more concrete, and easier to follow, we are going to proceed with $q = 2$.

The problem is now well defined mathematically:

$$\text{find a discount factor curve:} \quad df \in \mathcal{C}^{2,pl}\big([0,T],N\big) \text{ so that } \|v\|_p \text{ is minimized,}$$

where $\|v\|$ is the ℓ^p-norm of v,

$$v = (v_1, v_2, \ldots, v_I), \qquad v_i = \sum_{l=1}^{L_i} df(t_{il})\, p_{il};$$

v_i is the net present value of instrument 'i' given that the cash flow is discounted using the discount factors $df(t_{il})$. The EpiSolutions Inc. implementation relies on the ℓ^∞-norm,

$$\|v\|_\infty = \max\big[\,|v_1|, |v_2|, \ldots, |v_I|\,\big],$$

so let's proceed with this criterion but it should be noted that one can choose any $p \in [1,\infty)$ that might better represent the decision maker's preferences or concerns. In fact, except for extremely unusual portfolio, the differences between the solutions should be insignificant.

Since df belongs to $\mathcal{C}^{2,pl}([0,T],N$, it's of the form: for $k = 1,\ldots,N$, $\delta = T/N$, $t \in (\delta(k-1), \delta k]$ and $\tau = t - \delta(k-1)$

$$df(t) = 1 + v_0 t + \frac{1}{2}a_0 t^2 + \frac{\tau^3}{6}x_k + \sum_{j=1}^{k-1} x_j \left[\frac{\delta^3}{2}\left(\frac{1}{3} + (k-j-1)(k-j)\right) + \delta^2\tau(k-j-0.5) + \frac{\delta\tau^2}{2}\right],$$

where $a_0, v_0, x_1, x_2, \ldots, x_N$ are parameters to be determined; note that the discount factor at time $t = 0$ is 1, so we can fix this 'constant' (z_0). But simply being of this form does not make df a discount factor curve. We already have that $df(0) = 1$, we need to add two conditions:

- df should remain non-negative, thus we have to introduce the constraints: $df(t) \geqslant 0$ for all $t \in [0, T]$;
- df should be decreasing, at least non-increasing, this means that $df'(t) \leqslant 0$ for all $t \in [0, T]$, a condition that translates into the constraints:

$$df'(t) = v_0 + a_0 t + \sum_{j=1}^{k-1} x_j \big(\delta^2(k - j - 0.5) + \delta\tau\big) + \frac{\tau^2}{2} x_k \leqslant 0,$$
$$\forall t \in (0, T].$$

Putting this all together with df as defined above, yields the optimization problem:

$$\begin{aligned}
&\min \theta \\
&\text{so that } \theta \geqslant \sum_{l=1}^{L_i} df(t_{il})\, p_{il}, \quad i = 1, \ldots, I, \\
&\qquad \theta \geqslant -\sum_{l=1}^{L_i} df(t_{il})\, p_{il}, \quad i = 1, \ldots, I, \\
&\qquad df(t) \geqslant 0, \quad t \in [0, T], \\
&\qquad v_0 + a_0 t + \sum_{j=1}^{k-1} x_j \big(\delta^2(k - j - 0.5) + \delta\tau\big) + \frac{\tau^2}{2} x_k \leqslant 0, \quad t \in (0, T], \\
&\qquad v_0 \leqslant 0, \quad a_0 \geqslant 0, \quad x_k \in \mathbb{R}, \quad k = 1, \ldots, N;
\end{aligned}$$

the restriction $v_0 \leqslant 0$ means that $df'(0)$ is not positive, and $a_0 \geqslant 0$ says that the curve should have positive curvature at $t = 0$. The constraints involving θ tell us that

$$\theta \geqslant \max_{i=1,\ldots,I} \left[\left| \sum_{l=1}^{L_i} df(t_{il})\, p_{il} \right| \right],$$

and by minimizing θ, we minimize the max-error; this inequality is split into $2I$ constraints so that all constraints are linear.

We have a linear optimization problem with a finite number of variables $N + 3$, but with an infinite number of constraints ($\forall t \in [0, T]$). To solve this problem, one could consider using one of the techniques developed specifically for (linear) semi-infinite optimization problems. Because of the nature of the problem, however, one can safely replace the conditions involving 'for all $t \in [0, T]$' by for all $t \in \{1/M, 2/M, \ldots, T/M\}$ with M sufficiently large; in the EpiSolutions Inc. implementation M is usually chosen so that the mesh size $(1/M)$ is 1 month. After this time-discretization, the problem becomes a linear programming problem that can be solved using a variety of commercial

packages. In addition to the constraints described earlier, the version implemented at EpiSolutions Inc. also relies on a few additional constraints that will improve the shape the zero-curves to fit more specifically the context.

One important component of the *EpiCurves* solution is that all the zero-curves are defined at every time t, there is never any need to resort to interpolations to fill in missing time-gaps. This, of course, gives us great flexibility in choosing the right approximations when building pricing mechanisms.

1.6.3. More examples

Let's now consider and analyze a few more examples.

1.6.3.1. DFS-portfolio. Let's first go back to the DFS-example of Section 1.2. Of course, by just comparing the graphs of the forward rates curves provided by Financial Providers No. 1 and No. 2 and the corresponding EpiCurve, it's evident that there is the possibility that some financial factors/indicators might be a little bit too much off. Here is a specific example: On June 10, 1997, the June 1999 futures contract settled at \$93.36. After a small convexity adjustment, this implies a 90-day forward rate of approximately 6.60%. This contract was not included in the BootStrapping sub-portfolio (17 instruments) of Financial Provider No. 1. The forward rate supplied by Financial Provider No. 1 for this period is 5.10%! Not surprisingly, the forward rate for this period, supplied by *EpiCurves*, that takes into account *all* 36 instruments, is the more reliable 6.45%.

1.6.3.2. Bond-portfolio. The first one is a Bond-portfolio. This data-set includes U.S. Treasury Bill and U.S. Treasury Bond data from August 3, 2001. There are 7 instruments in all, including 3 U.S. Treasury Bills with term-to-maturities ranging from 3 to 7 months and 4 U.S. Treasury Bonds with term-to-maturities ranging from 2 to 30 years. This data was obtained from the Bloomberg U.S. Treasuries web page; details are in Table A.4 of Appendix A. As a point of comparison, we use the results of the BootStrapping technique supplied by Financial Provider No. 2; Financial Provider No. 1 BootStrapping functionality can not deal with a Bond-portfolio. The results are graphed in Figure 5.

1.6.3.3. DFS2-portfolio. This next example is a relatively challenging one. The portfolio includes 51 instruments: Deposits, Futures and Swaps from August 3, 2001 with quite a bit of overlap of maturity-dates. A short description of the composition of this portfolio follows here; details are in Tables A.5, A.6 of Appendix A. There are 51 instruments in all, broken down as follows: 3 Eurodollar Deposits with term-to-maturities ranging from 1 to 6 months, 40 Eurodollar Futures with 90-day deposit maturities ranging from 4 months to 10 years, and 8 Swaps with term-to-maturities ranging from 1 to 10 years. This data was obtained from the Federal Reserve (Statistical Release H.15) and the Chicago Mercantile Exchange (CME). The results are graphed in Figures 6 and 7.

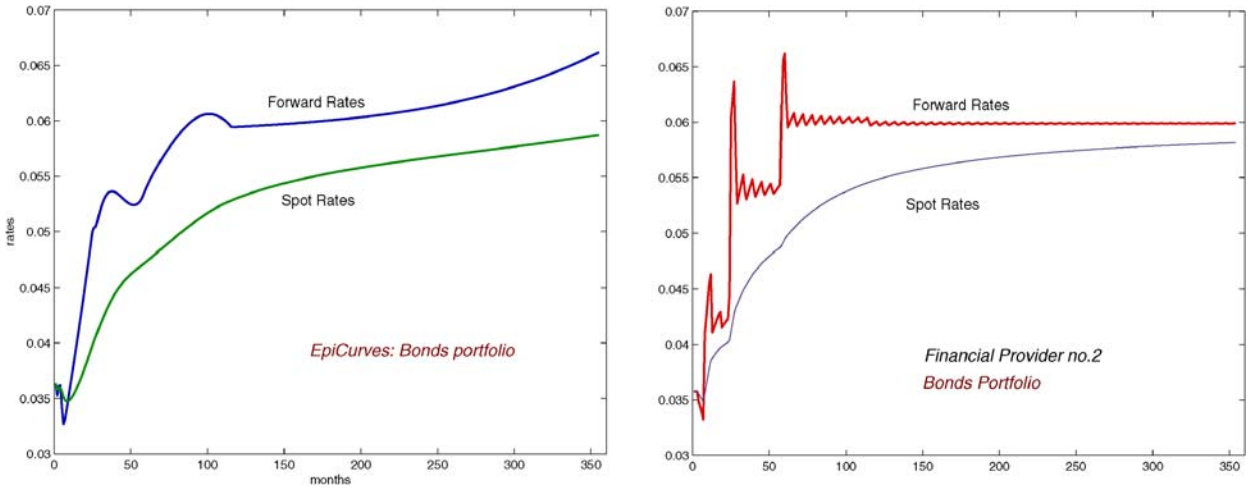


Fig. 5. Spot and Forward Rates associated with a Bonds portfolio.

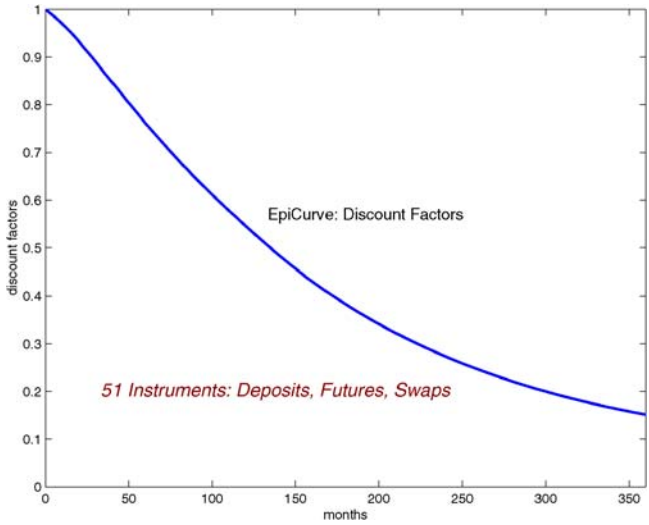


Fig. 6. Discount Factor curve for the 51-instruments portfolio.

In the EpiSolutions Inc. implementation of the *EpiCurves* methodology, there is an option that allows the user to fine tune the level of accuracy that will be acceptable; accuracy being defined in terms of the max-error, i.e., in terms of the objective of the optimization problem. Asking for a higher level of accuracy will usually result in a more jagged curve since one must accommodate/adjust more rapidly to even small changes in the cash flow. This is effectively illustrated by curves graphed in Figure 7. In the first one the tolerance is 5 base points, in the second one just 1 base point.

Notwithstanding this fine tuning, *EpiCurves* is really the only methodology that provides 'serious' zero-curves associated with such a portfolio. The only BootStrapping approach that one could rely on, if one can use the word 'rely', is the one implemented by EpiSolutions Inc. But the results are less than satisfactory, see Figure 8. Both the spot and the forward rate curves were derived for this portfolio with the forward rate curve generated by BootStrapping spiking up to 30% at one point and then immediately thereafter going negative! This suggested massaging the portfolio by introducing a *convexity adjustment* and, indeed, this improves substantially the BootStrapping results although the forward rate curve comes with some abrupt rate changes; note that this convexity adjustment has only a minor effect on the *EpiCurves* results.

The major objective in developing the *EpiCurves* methodology was to overcome the inconsistent assumptions and limitations of the standard BootStrapping technique and its Maximum Smoothness variant. This is accomplished by allowing for inclusion of the *complete portfolio* of term structure instruments, while at the same time providing the *smoothness* so crucial to practitioners as a solid foundation on which to build believable valuations, forecasts and other financial analytics.

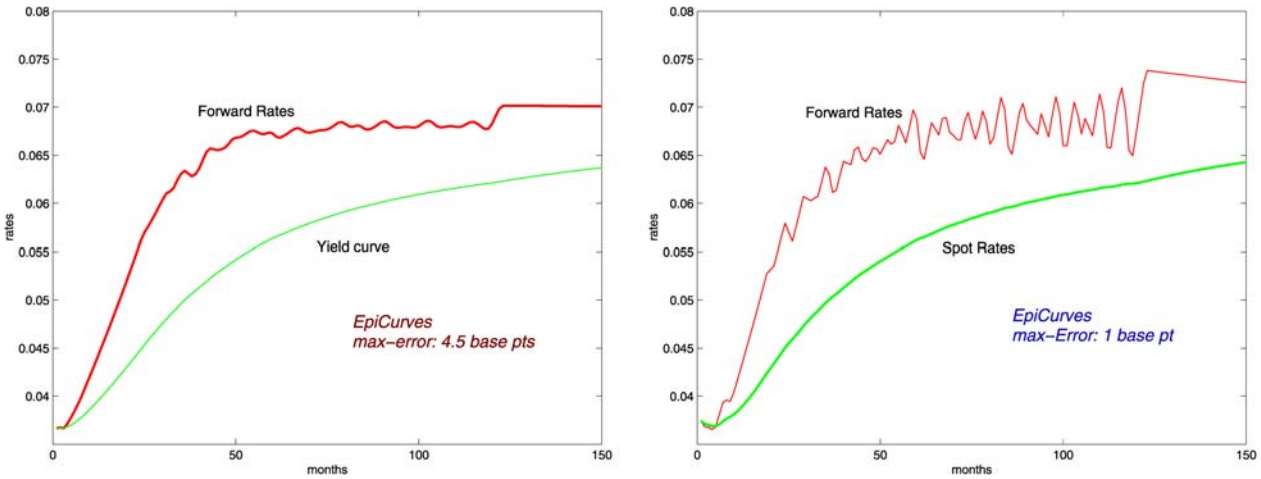


Fig. 7. Variations of the zero-curves under max-error tolerance.

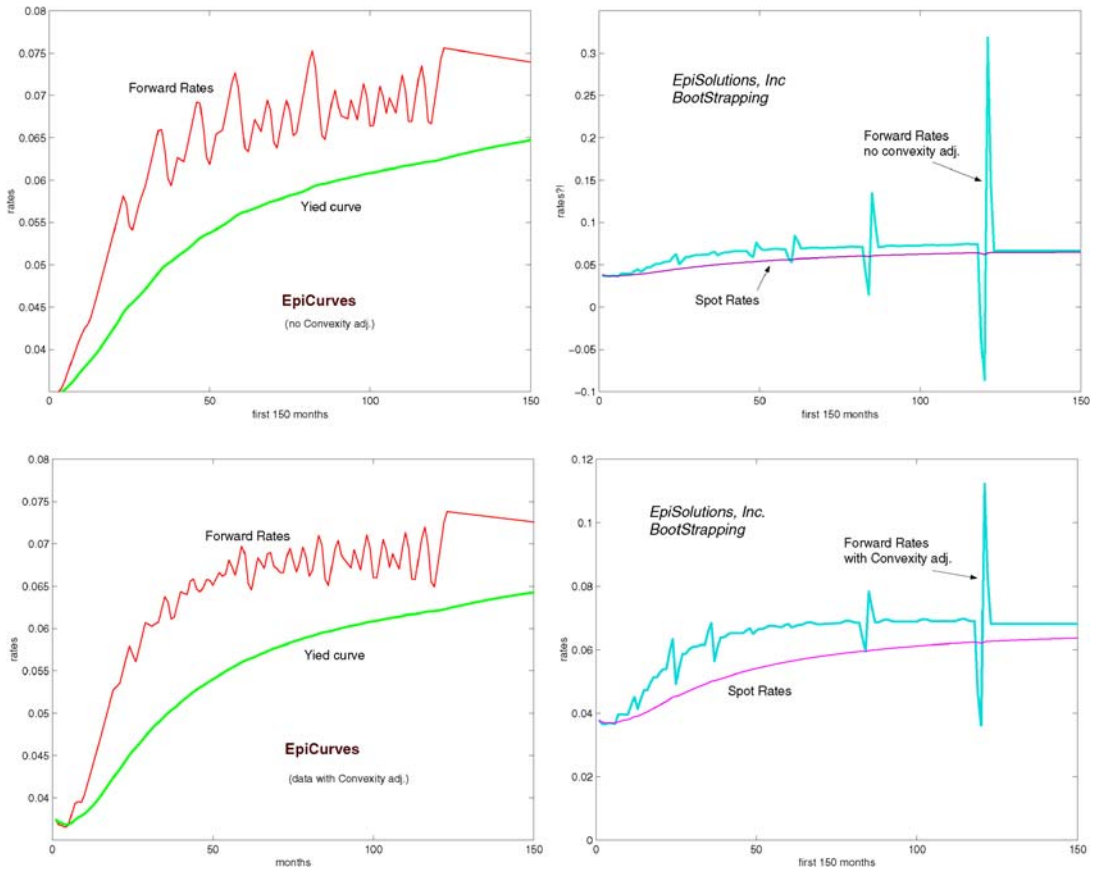


Fig. 8. Spot and Forward Rates with and without Convexity adjustments.

1.7. A comparison for U.S. Treasury curves

In this section we perform an empirical analysis of several different estimation methods using monthly U.S. Treasury bond data obtained from Mergent's Bond Record. The data set covers the period from January 31, 1999 through December 31, 2003, for a total of 60 observations (with each observation, or portfolio, containing approximately 100–110 bonds). In particular, we focus on the following methods:

- *EpiCurves, run for accuracy*: With this approach we run EpiCurves with the objective of minimizing the maximum absolute pricing error.
- *EpiCurves, run for smoothness*: With this approach we run EpiCurves with the objective of 'maximizing' the smoothness of the resulting curves. We employ Van Deventer's smoothness criterion, divided by T, to facilitate comparison of the monthly bond portfolios. We then iterate, relaxing the maximum pricing error constraint until we have achieved an acceptable level of smoothness (defined to be 0.01, for this analysis).
- *Van Deventer's maximum smoothness approach*: With this approach the primary control variable is the number of spline segments. Too many segments may result in more accurate pricing, with a resulting loss in smoothness, while too few may result in a spline that is too stiff, with a resulting loss in pricing accuracy. To address this, we set the initial number of spline segments to be approximately equal to the square root of the number of bonds in a given portfolio. We then iterated, reducing the number of spline segments by one, until the number of segments giving the best pricing results was found.
- *Svensson's extension of the Nelson–Siegel model*: With this approach we used a number of different sets of starting values, solved for each, and selected the solution with the best pricing results. β_0 and β_1 were held constant, with β_0 set to the yield-to-maturity of the bond with the longest maturity in a given portfolio, and β_1 set to the difference between this and the yield-to-maturity of the bond with the shortest maturity in the portfolio. This left us with 4 parameters (β_2, β_3, τ_1, τ_2) each of which we allowed 5 possible values, giving us a total of 625 sets of starting values.

We examined the estimation results along several different dimensions:

- *Smoothness*: Van Deventer's smoothness criterion, divided by T.
- *In-Sample Mean Squared Error* (*IN MSE*): the mean squared pricing error of bonds used to perform the estimation.
- *In-Sample Weighted-Average Absolute Error* (*IN WAE*): the weighted-average absolute pricing error of bonds used to perform the estimation, weighted by $1/\mathit{maturity}$.
- *In-Sample Maximum Absolute Error* (*IN MAX*): the maximum absolute pricing error of bonds used to perform the estimation.
- *Out-of-Sample Mean Squared Error* (*OUT MSE*): the mean squared pricing error of the bonds *not* used to perform the estimation.
- *Out-of-Sample Weighted-Average Absolute Error* (*OUT WAE*): the weighted-average absolute pricing error of bonds *not* used to perform the estimation, weighted by $1/\mathit{maturity}$.

- *Out-of-Sample Maximum Absolute Error* (*OUT MAX*): the maximum absolute pricing error of bonds *not* used to perform the estimation.
- *Speed*: the number of minutes to complete an estimation on a 2.0 GHz machine running Windows 2000.

All pricing error results are given in basis points. *In-sample* bonds were defined by starting with shortest maturity bond in a given portfolio, and including every other bond in ascending maturity order—with the caveat that the longest maturity bond was always *in-sample*. We did not filter the data in any other way, nor did we employ any outlier exclusion scheme during the estimations. It should be noted, however, that there were no Treasury bills or callable bonds in the data set. Some graphical estimation results can be seen in Figure 9. The numerical results are summarized in Table 1.

We can make some general observations about the results:

- The Van Deventer and Svensson results are always smoother than the EpiCurves results. This is not surprising, since they are *defined* to be so, while EpiCurves allows the user to trade-off important criteria.
- The Van Deventer method performs well in upward-sloping term structure environments (2002–2003), and poorly when the term structure is flatter (1999–2001). This is also not very surprising, since fitting splines to straight lines is difficult.
- Over the entire period studied, the Van Deventer results where more variable than the results from any of the other methods considered.
- The Svensson results were remarkably stable, while providing relatively good pricing results, in different environments.
- When run for *accuracy*, EpiCurves always gave the best pricing results.
- When run for *smoothness*, EpiCurves was able to achieve an comparable level of smoothness to the other methods, while retaining excellent pricing accuracy.
- EpiCurves allowed for effective trade-off between smoothness (for strategic decision making) and pricing accuracy (for tactical decision making).
- EpiCurves was always *faster* than Van Deventer and Svensson, by a wide margin.

The last point deserves further comment. As can be seen in Table 1, even in the worst case, EpiCurves performed the estimations 4 times faster than either Van Deventer or Svensson. The original Van Deventer method was specified using a small set of zero-coupon bond prices, which requires solving a system of linear equations *once*. In the extension presented here, we are using large number of coupon-bearing bonds, making it necessary to iterate in two dimensions. In the first dimension, given a set of knot points for the spline segments (also known as term structure vertexes), we iterate to find the set of zero-coupon bond prices for the knot points that minimizes the overall pricing error. In the second dimension, we also iterated over the number of knot points to determine the optimal trade-off between curve flexibility and pricing accuracy. In the best case, this required an average of 27 minutes per estimation (2003). In the worst case, the speed deteriorated to an average of 46 minutes per estimation (1999).

One of the main criticisms of the Svensson methodology is that the parameters can be difficult to estimate. This is due to the fact that the spot and forward rate functions, though linear in the β's, are non-linear the τ's. As a result, there are multiple local

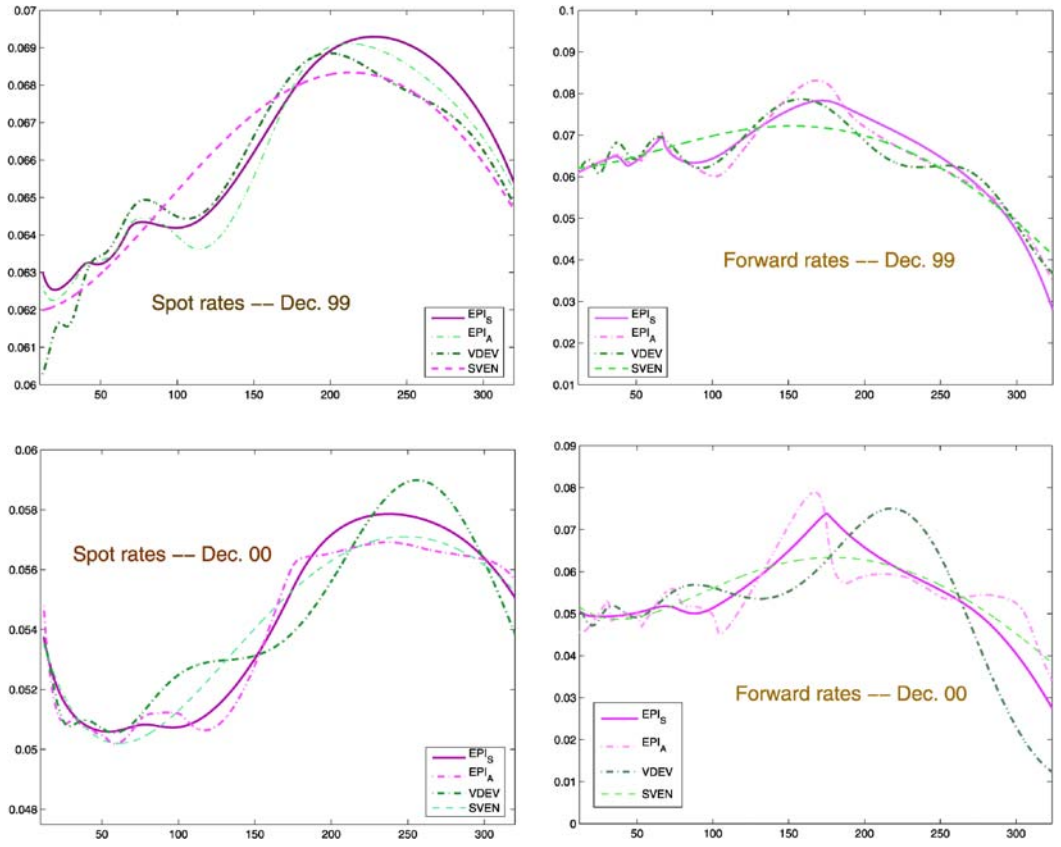


Fig. 9. Comparison of Spot and Forward Rate curves.

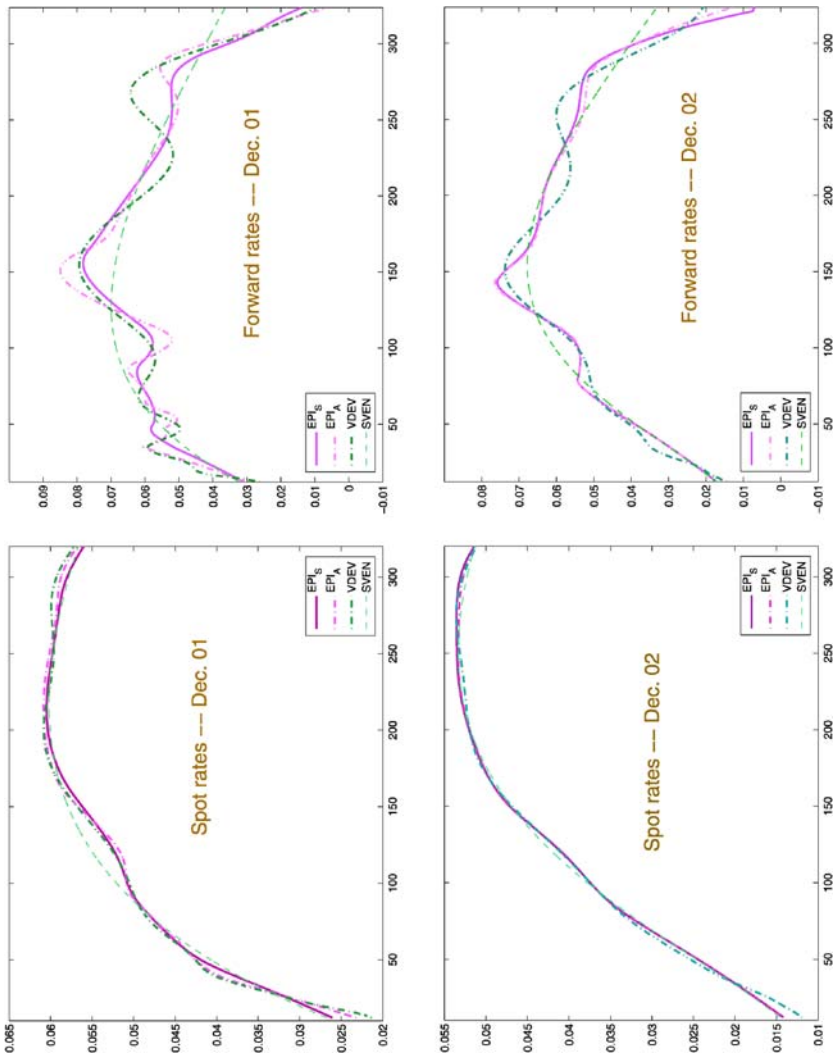


Fig. 9. *(Continued.)*

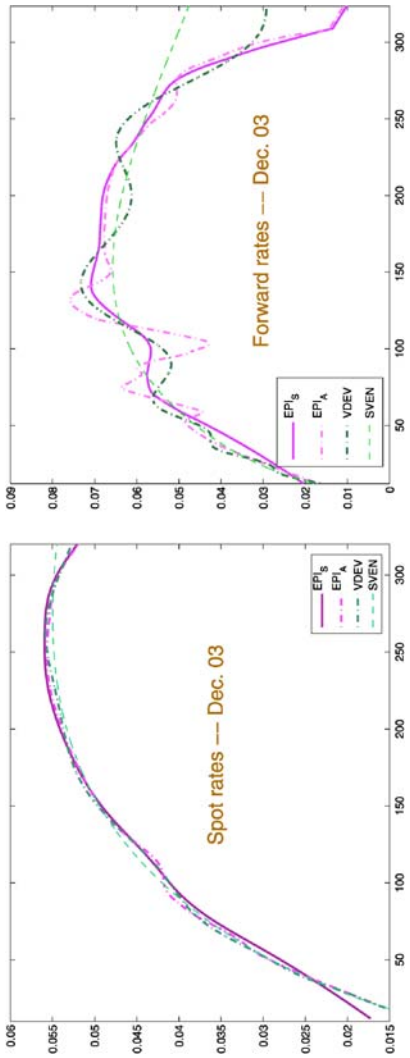


Fig. 9. *(Continued.)*

Table 1
U.S. Treasury curve statistics (averages)

1999–2003	EPI accurate	EPI smooth	Van Deventer	Svensson
Smoothness	0.252989	0.005163	0.000282	0.000877
IN MSE	442.03	986.59	2230.32	1214.64
IN WAE	8.95	22.42	8.66	20.74
IN MAX	58.44	89.16	145.03	133.38
OUT MSE	501.41	1032.69	2075.52	1040.42
OUT WAE	9.63	22.50	8.86	20.52
OUT MAX	79.54	96.82	151.85	122.77
Speed (min)	2	4	34	30
2003	EPI accurate	EPI smooth	Van Deventer	Svensson
Smoothness	0.731324	0.003204	0.000028	0.000559
IN MSE	298.25	1378.76	448.76	1647.35
IN WAE	4.51	34.83	5.72	22.80
IN MAX	41.86	97.72	80.70	128.69
OUT MSE	321.82	1291.56	323.74	1496.99
OUT WAE	5.87	34.55	5.89	22.53
OUT MAX	58.23	80.86	72.51	121.62
Speed (min)	4	6	27	25
2002	EPI accurate	EPI smooth	Van Deventer	Svensson
Smoothness	0.357384	0.004697	0.000112	0.003827
IN MSE	568.15	1390.76	604.60	1549.93
IN WAE	6.42	32.13	6.34	29.42
IN MAX	57.40	111.04	91.89	142.65
OUT MSE	637.45	1463.28	615.65	1391.80
OUT WAE	7.55	32.33	6.65	29.42
OUT MAX	85.89	116.09	105.80	152.67
Speed (min)	3	6	30	26
2001	EPI accurate	EPI smooth	Van Deventer	Svensson
Smoothness	0.103716	0.003786	0.000104	0.000000
IN MSE	215.98	530.63	2738.24	1072.05
IN WAE	8.18	15.11	10.36	26.68
IN MAX	39.54	61.68	151.60	110.96
OUT MSE	355.28	693.17	2376.87	973.28
OUT WAE	9.27	16.11	10.74	26.77
OUT MAX	72.58	91.73	172.03	107.12
Speed (min)	2	4	30	28

minima, making it necessary to run the estimations for many different sets of starting values. To completely eliminate uncertainty in the results would require running the estimations over an unwieldy number of sets of starting values. By holding β_0 and β_1

Table 1
(*Continued*)

2000	EPI accurate	EPI smooth	Van Deventer	Svensson
Smoothness	0.02769	0.006766	0.000098	0.000000
IN MSE	300.23	539.01	3114.29	623.21
IN WAE	7.47	8.69	9.29	10.99
IN MAX	55.75	67.35	162.62	118.72
OUT MSE	356.57	616.53	2900.21	408.48
OUT WAE	7.63	8.72	9.55	10.46
OUT MAX	74.62	82.71	165.51	81.49
Speed (min)	1	3	34	33
1999	EPI accurate	EPI smooth	Van Deventer	Svensson
Smoothness	0.044831	0.007363	0.001069	0.000000
IN MSE	827.53	1093.78	4245.72	1180.67
IN WAE	18.19	21.35	11.59	13.82
IN MAX	97.65	108.00	238.32	165.90
OUT MSE	835.93	1098.92	4161.15	931.56
OUT WAE	17.82	20.81	11.49	13.43
OUT MAX	106.38	112.73	243.38	150.98
Speed (min)	2	3	46	38

constant, we reduced the number of free variables to 4. Then allowing these 4 variables to take on each of 5 different values, we settled on 5^4, or 625, sets of starting values. The β's are ranged from $-$(maximum bond yield) to $+$(maximum bond yield), and the τ's are ranged from minimum bond maturity to maximum bond maturity. In the best case, this required an average of 25 minutes per estimation (2003). In the worst case, the speed deteriorated to an average of 38 minutes per estimation (1999).

2. Volatility structure

2.1. Setting the stage

The motivation for the *EpiVolatility* model is to provide a consistent, flexible, and *market calibrated* term structure of volatility that, in particular can serve as input to any valuation package, in particular the *EpiValuation* library of EpiSolutions Inc. No comprehensive comparisons will be made with alternative approaches because, to our knowledge, the alternative methods that have been suggested are either proprietary (Risk Metrics), are based on historical data rather than (present-day) market data, are relatively elementary (regression, linear interpolation, for example) or rely on a Boot-Strapping type approach that will be described below; a brief review of this literature

was provided by Dupačová and Bertocchi (2001, §3) see also an earlier article by these authors and their collaborators (Dupačová et al., 1997).

The primary option pricing models underlying a significant number of valuation packages, including the EpiValuation Library, are the Black, Derman and Toy (BDT) (1990) binomial model and the Black (1976) model for interest-rate derivatives. The approach outlined in this contribution is in the class of popular interest rate models known as market models. In particular, our estimation of the term structure of volatility is based on the standard market model, which is also the basis for volatility estimation in the LIBOR market model (LMM). This makes *EpiVolatility* model consistent with the LIBOR market model.

The LIBOR market model, also known as the Brace, Gatarek and Musiela (BGM) (1997) model, is an extension of the Heath, Jarrow and Morton (HJM) (1992) model. However, where the HJM model is based on (unobservable) instantaneous forward rates, the LMM is based on observable market LIBOR rates that follow a lognormal process. This makes the LMM consistent with the Black model for pricing interest rate caps and floors, which is used by market practitioners. A similar model was developed by Jamshidian (1997) for swap rates that is consistent with the Black model for valuing European swaptions. This model is known as the Swap market model (SMM). One can refer to Hull (2000) for a detailed description of these market and valuation models.

Although the LMM and the SMM are each internally consistent (neither allows opportunities for arbitrage), they are not consistent with each other, see Hull and White (2000). This is because the LMM is based on a lognormal process for forward rates and the SMM is based on a lognormal process for swap rates, where swap rates can be thought of as an average of a series of forward rates. However, the difference in swaption prices between the two models is low, see Brigo, Capitani and Mercurio (2001), and that the SMM substantially overprices caplets, see De Jong, Driessen and Pelsser (2001). We want to be clear in stating, therefore, that for practical purposes one can reasonably *assume* that the LMM is the preferable model.

The construction of a BDT binomial interest rate tree requires three inputs: a time ruler, a yield curve, and a volatility curve. The time ruler is based on the security being priced (timing of cash flows, options, etc.). In Section 1, we have dealt with the construction of the LIBOR yield curve (or zero-curve) as estimated from current market rates/prices of Eurodollar deposits, Eurodollar futures, and/or (on market) interest rate swaps. In principle, the volatility curve (or term structure of volatility) can simply be observed in the market, since interest rate cap and floor prices are *quoted* in terms of flat implied Black volatilities.

Table 2 has the cap volatility quotes from GovPx on October 24, 2002. The term of the cap is expressed in years and the tenor of the cap is 3 months (i.e., each cap is a series of 3 month caplets). The bid and ask volatilities are expressed in percent per annum. The strike is the at-the-money strike rate of the cap. This means that the cap strike rate equals the swap rate, for a swap with the same reset dates as the cap.

Caps, however, are quoted in terms of *flat* volatilities. A flat volatility is the implied volatility of the cap, when that volatility is applied to *all* the caplets underlying the cap.

Table 2
Cap volatility quotes

Term	Bid vol	Ask vol	Strike
0.50	52.86	53.86	1.722
1.00	50.20	51.20	1.894
1.50	49.90	50.90	2.201
2.00	48.26	49.26	2.533
2.50	45.09	46.09	2.826
3.00	42.28	43.28	3.073
3.50	40.13	41.13	3.287
4.00	37.90	38.90	3.481
4.50	36.18	37.18	3.628
5.00	34.51	35.51	3.769
6.00	31.65	32.65	4.046
7.00	29.72	30.72	4.265
8.00	28.00	29.00	4.456
9.00	26.70	27.70	4.594
10.00	25.47	26.47	4.730

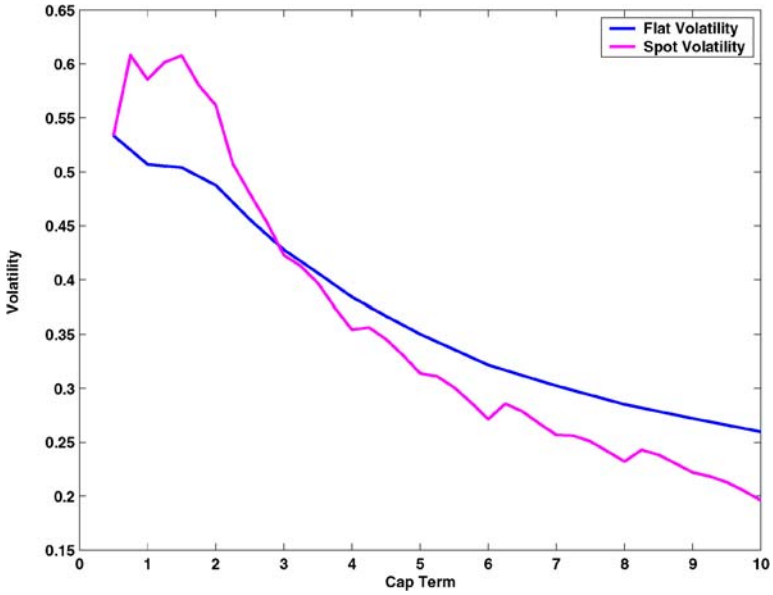


Fig. 10. Flat and spot volatility curves.

As a result, caplets underlying more than one cap (for example, the 9×12 caplet is common to all caps in the table above, except the 6 month cap), will be priced with different volatilities depending on the cap being considered. An alternative approach is

to use a unique volatility for each caplet in the cap series (i.e., for the 9 × 12 caplet, a single volatility would be used regardless of the cap being valued). These are called *spot* volatilities. Spot volatilities can be deduced from flat volatilities using a standard BootStrapping approach. With this approach, a series of cap prices is first generated using the flat volatilities, then the difference between each cap price and the previous cap price gives a forward caplet price, the Black model can then be inverted to produce the implied volatility for this caplet. This is the spot volatility of the forward rate with the same term as the caplet. Figure 10 shows the flat volatilities and spot volatilities [obtained by BootStrapping] from Table 2, based on mid-market quotes.

For some applications, derivation of spot volatilities in this fashion may be acceptable. But there are some serious limitations to the way this data may be used. We will discuss these limitations and how the EpiVolatility model attempts to overcome them, but first the valuation of caps and floors with the Black model, and the Black, Derman and Toy interest rate model are briefly discussed.

2.2. *Some tree-based valuation models*

An interest rate cap (floor) is simply a portfolio of European call (put) options on forward interest rates.

Consider a caplet with term δ_n, notional value L, and strike rate R_K. Let $F_{t_n,t_{n+1}}$ be the interest rate for the period between time t_n and time t_{n+1} (where $\delta_n = t_{n+1} - t_n$), observed at time t_n. The caplet has a payoff at time t_{n+1} of

$$L\delta_n(F_{t_n,t_{n+1}} - R_K)^+. \tag{1}$$

If the rate $F_{t_n,t_{n+1}}$ follows a lognormal process with volatility σ, the familiar Black pricing formula can be used to determine the value of the caplet at time $t_0 < t_n$ as

$$C_n = L\delta_n P(t_0, t_{n+1})\big(F_{t_n,t_{n+1}} N(d_1) - R_K N(d_2)\big) \tag{2}$$

where

$$d_1 = \frac{\ln(F_{t_n,t_{n+1}}/R_K) + \sigma^2 t_n/2}{\sigma\sqrt{t_n}}, \qquad d_2 = \frac{\ln(F_{t_n,t_{n+1}}/R_K) - \sigma^2 t_n/2}{\sigma\sqrt{t_n}}$$

and $N(\cdot)$ is the cumulative Gaussian distribution with mean 0 and variance 1. $P(t_0, t_{n+1})$ is the price, at time t_0, of a zero coupon bond maturing at time t_{n+1}.

The formula for the corresponding floorlet is

$$F_n = L\delta_n P(t_0, t_{n+1})\big(R_K N(-d_2) - F_{t_n,t_{n+1}} N(-d_1)\big). \tag{3}$$

Since a cap (floor) is a portfolio of caplets (floorlets), the price of a cap with term $T = t_{n+1}$ at time $t_0 < t_n$ is

$$C = \sum_{i=1}^{n} C_n \quad \text{and} \tag{4}$$

$$F = \sum_{i=1}^{n} F_n \tag{5}$$

for a floor.

In 1990, Black, Derman and Toy proposed an algorithm for constructing a binomial interest rate tree that yields a discretized version of the model:

$$\mathrm{d}\ln r(t) = \left[\theta(t) + \frac{\sigma'(t)}{\sigma(t)}\ln(r)\right]\mathrm{d}t + \sigma(t)\,\mathrm{d}z \tag{6}$$

where $\sigma'(t)$ is the partial derivative of σ with respect to t, and $\theta(t)$ is a time-dependent parameter used to fit the model to the initial term structure (Hull, 2000). Therefore, in order to use this model the volatility function must be differentiable at any time t.

2.3. The EpiVolatility model

2.3.1. Guidelines

The following are the desirable properties of a volatility curve, for use as a basis for interest rate modeling and security valuation (present and future):

- The partial derivative with respect to t, at any time t, exists and is continuous.
- The curve is defined with respect to *spot* (not flat) volatility, also known as *forward–forward* volatility.
- Spot volatility may be retrieved for a term of *arbitrary* length.
- A particular form for the volatility function need *not* be assumed.

The first three properties are requirements for the successful implementation of the BDT model in either binomial tree or Monte Carlo form. The last property is desirable from an implementation perspective, since we are not forced to assume a particular functional 'shape' that may or may not be representative of current market conditions. Further, any assumed functional form will restrict us to a few parameters for fitting the curve to the market data.

Figure 10 shows that the BootStrapped spot volatility curve does not meet these criteria:

- The partial derivative with respect to t, at any time t, is not continuous, and might even fail to exist,
- The forward volatilities retrieved from this curve are only defined for forward rates with the same start and end times as the caplets used in its construction.

Regarding this second point, if the tenor of the caps in the market data set is 3 months (for example), then the volatility curve only contains information about the forward volatilities of 3 month LIBOR forward rates that coincide with caplet dates. The curve says nothing about forward volatilities of 3 month LIBOR forward rates that do not coincide with caplet dates, nor about forward volatilities of forward rates with any other term (1 day, 1 week, 1 month, etc.). But when constructing a BDT tree for arbitrary δt, this is exactly what is needed.

2.3.2. EpiSolutions Inc.'s approach

Here is a brief outline of an approach to volatility estimation, that is designed to meet the *desirable* criteria listed above:

- Start with market data (broker quotes) for interest rate caps or floors.
- Fit a (smooth) curve to the flat volatility quotes (mid, bid, or ask), using *EpiCurves*, see Section 1.6.
- Fit a (smooth) curve to the strike rate curve, using *EpiCurves*.
- Construct a set of caps or floors with terms every n months (where n depends on the tenor of the securities in the market data set, typically 3 or 6 months). The price for each security is determined using the Black model, and flat volatilities and strike rates read from their respective (smooth) curves.
- Fit a (smooth) curve to the cap or floor prices determined in the previous step using *EpiCurves*.

This last curve is the end product of the *EpiVolatility* model, and is used as the input to the *EpiValuation* library. When the *EpiValuation* library needs a spot volatility for a given term, it derives a cap or floor price for that term from the price curve, and then inverts the Black model to retrieve the corresponding *implied* spot volatility.

Here again, the meaning of 'smooth' is bound to be mostly subjective. One could take as definition, the number of times the curve is continuously differentiable. But there are analytic curves, i.e., of class $\mathcal{C}^\infty$, that do not 'look' smooth; refer to Sections 1.4, 1.6 for more about this issue. Practically, we shall content ourselves with curves that are less than infinitely smooth, but where we control the rate of change of the 2nd, or higher, derivative.

In the next section, we discuss the EpiSolutions criterion and methodology for fitting 'smooth' curves to market data. Finally, we give an example of the implementation of the *EpiVolatility* model.

2.4. Implementation

Using the cap price data from Tables 2, 3 shows the results of the first three steps outlined in Section 2.4

- Fit a (smooth) curve to the flat volatility quotes (mid, bid, or ask), using *EpiCurves*, cf. Section 1.6.
- Fit a (smooth) curve to the strike rate curve, using *EpiCurves*.
- Construct a set of caps or floors with terms every n months (where n depends on the tenor of the securities in the market data set, typically 3 or 6 months). The price for each security is determined using the Black model, and flat volatilities and strike rates read from their respective (smooth) curves.

  We use mid market volatility quotes and the cap tenors are assumed to be 3 months.

The last step is to fit a curve to the price vector from Table 3, using the *EpiCurves* technology. The result can be seen in Figure 11.

Table 3
Calibrating cap set

Term	Flat vol	Strike	Price	Term	Flat vol	Strike	Price
0.50	53.3539	1.7222	0.000595	5.50	33.4571	3.9121	0.053636
0.75	51.9915	1.7855	0.001613	5.75	32.7672	3.9811	0.056743
1.00	50.6982	1.8943	0.002945	6.00	32.1487	4.0461	0.059799
1.25	50.3926	2.0381	0.004590	6.25	31.6054	4.1060	0.062867
1.50	50.3984	2.2025	0.006613	6.50	31.1180	4.1615	0.066044
1.75	49.8694	2.3706	0.008936	6.75	30.6629	4.2141	0.069153
2.00	48.7541	2.5337	0.011497	7.00	30.2180	4.2652	0.072157
2.25	47.2321	2.6858	0.014139	7.25	29.7684	4.3160	0.075074
2.50	45.5900	2.8260	0.016878	7.50	29.3227	4.3657	0.078018
2.75	44.0749	2.9551	0.019636	7.75	28.8948	4.4129	0.080902
3.00	42.7722	3.0737	0.022441	8.00	28.4975	4.4563	0.083782
3.25	41.6809	3.1831	0.025259	8.25	28.1391	4.4948	0.086745
3.50	40.6276	3.2872	0.028185	8.50	27.8108	4.5296	0.089836
3.75	39.4827	3.3890	0.031055	8.75	27.5007	4.5622	0.092862
4.00	38.3912	3.4817	0.034002	9.00	27.1973	4.5943	0.095746
4.25	37.4882	3.5592	0.037239	9.25	26.8918	4.6271	0.098460
4.50	36.6764	3.6283	0.040725	9.50	26.5841	4.6608	0.101054
4.75	35.8452	3.6978	0.044101	9.75	26.2762	4.6952	0.103437
5.00	35.0100	3.7690	0.047326	10.00	25.9700	4.7300	0.105574
5.25	34.2084	3.8409	0.050472				

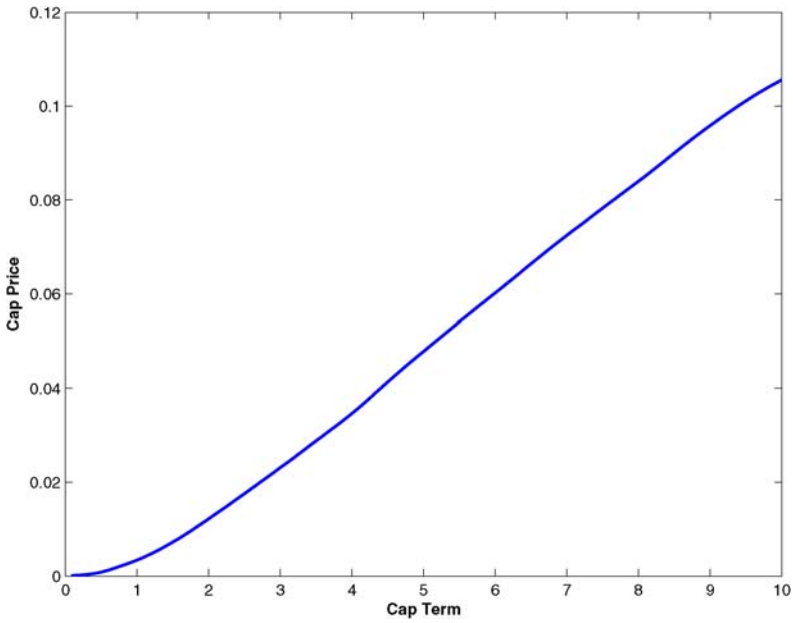


Fig. 11. Fitted cap price curve.

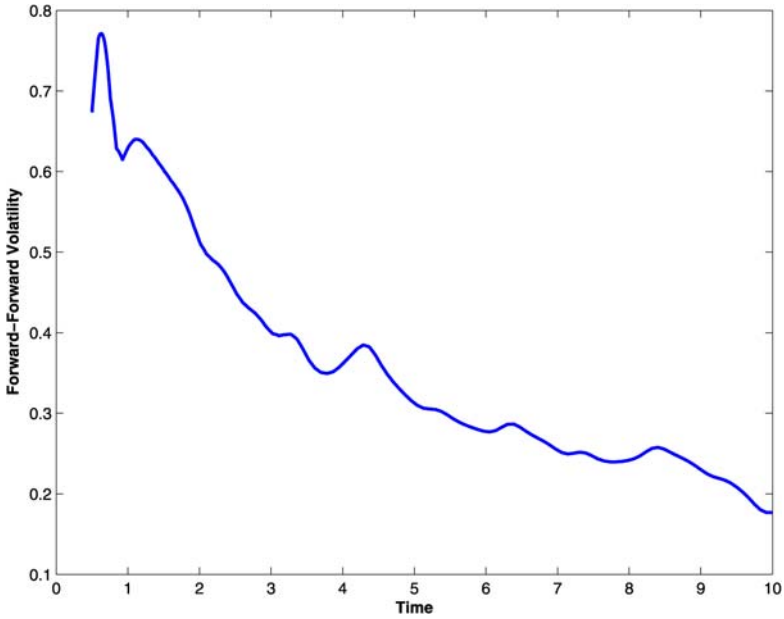


Fig. 12. Daily spot volatilities (3 month rates).

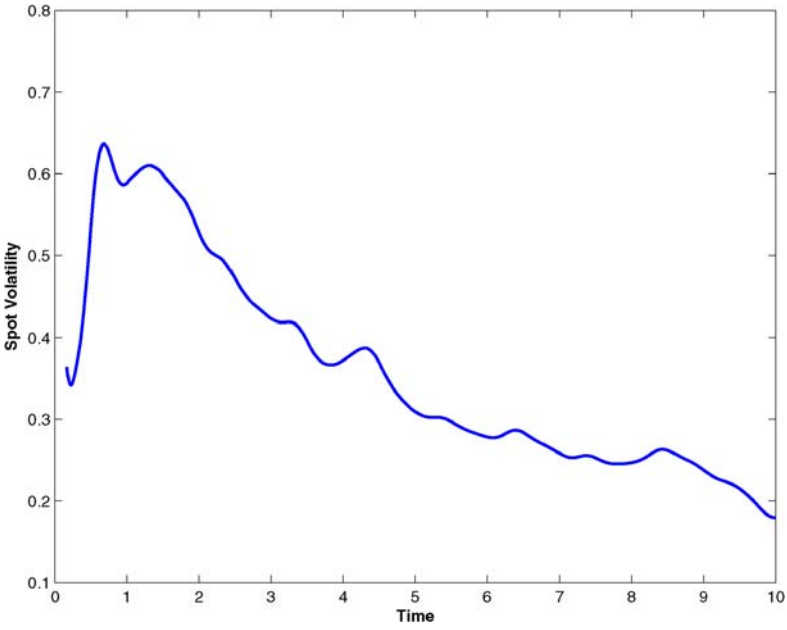


Fig. 13. Daily spot volatilities (1 month rates).

This fitted cap price curve is then used as input to the *EpiValuation* library, which uses it to determine spot volatilities in interest rate modeling and option pricing. The steps the *EpiValuation* library uses to do this are straight-forward. Given a time step from time t_n to time t_{n+1},

- Read a cap price C_{t_n} from the curve with term t_n.
- Read a cap price $C_{t_{n+1}}$ from the curve with term t_{n+1}.
- The forward caplet price $C_{t_n,t_{n+1}}$ then simply equals $C_{t_{n+1}} - C_{t_n}$.
- Invert the Black model for the forward caplet (i.e., solve for volatility instead of price), assuming the caplet is at-the-money. This means the strike rate is set equal to the (simply compounded) forward rate $F_{t_n,t_{n+1}}$ for the same period.

Figure 12 shows the results for daily spot volatilities for 3 month Libor rates. This means that a volatility read from the curve is the volatility for the 3 month Libor rate starting at that time and maturing 3 months later. Figure 13 shows the results for 1 month Libor Rates.

2.5. Summary

The approach to estimating LIBOR spot volatilities outlined in Section 2 of this chapter arose from a need for a consistent and flexible method for determining quantities required by the *EpiValuation* library. The *EpiValuation* library is used for a number of functions, including:

- Current valuation of securities and portfolios.
- Future valuation of securities and portfolios.
- Scenario based total return estimation for securities and portfolios.

And as indicated in the Introduction, scenario-based total returns are themselves input for our *EpiManager*, that deals with portfolio optimization by relying on stochastic programming techniques.

Appendix A

Table A.1
Eurodollar deposits

Settle	Maturity	Rate
06/10/97	06/11/97	0.054688
06/11/97	07/11/97	0.056250
06/11/97	08/11/97	0.057500
06/11/97	09/11/97	0.057344
06/11/97	12/11/97	0.058125
06/11/97	06/11/98	0.061875

Table A.2
Swaps

Settle	Maturity	Rate
06/11/97	06/11/99	0.062010
06/11/97	06/11/00	0.064320
06/11/97	06/11/01	0.065295
06/11/97	06/11/02	0.066100
06/11/97	06/11/04	0.067860
06/11/97	06/11/07	0.068575

Table A.3
Eurodollar futures

Delivery	Maturity	Price	Delivery	Maturity	Price
06/16/97	09/16/97	94.195	06/19/00	09/19/00	93.220
09/15/97	12/15/97	94.040	09/18/00	12/18/00	93.190
12/15/97	03/16/98	93.820	12/18/00	03/19/01	93.190
03/16/98	06/16/98	93.725	03/19/01	06/19/01	93.120
06/15/98	09/15/98	93.610	06/18/01	09/18/01	93.080
09/14/98	12/14/98	93.510	09/17/01	12/17/01	93.050
12/14/98	03/15/99	93.410	12/17/01	03/18/02	92.980
03/15/99	06/15/99	93.390	03/18/02	06/18/02	92.980
06/14/99	09/14/99	93.360	06/17/02	09/17/02	92.940
09/13/99	12/13/99	93.330	09/16/02	12/16/02	92.900
12/13/99	03/13/00	93.260	12/16/02	03/17/03	92.830
03/13/00	06/13/00	93.250	03/17/03	06/17/03	92.830

Table A.4
Bond portfolio—U.S. Treasury bills and bonds

Settle	Maturity	Price	Settle	Maturity	Coupon	Price
08/03/01	11/01/01	3.44	08/03/01	07/31/03	0.03875	$99 + 30/32$
08/03/01	01/31/02	3.36	08/03/01	05/15/06	0.04625	$99 + 26/32$
08/03/01	02/28/02	3.33	08/03/01	02/15/11	0.05000	$98 + 25/32$
			08/03/01	02/15/31	0.05375	$99 + 00/32$

Table A.5
51-Instruments portfolio—Eurodollar futures

Delivery	Maturity	Price	Convex.	Delivery	Maturity	Price	Convex.
09/17/01	12/17/01	96.430	0.023	09/18/06	12/18/06	93.340	13.999
12/17/01	03/18/02	96.295	0.117	12/18/06	03/19/07	93.250	15.364
03/18/02	06/18/02	96.150	0.274	03/19/07	06/19/07	93.290	16.794
06/17/02	09/17/02	95.805	0.494	06/18/07	09/18/07	93.260	18.283
09/16/02	12/16/02	95.420	0.775	09/17/07	12/17/07	93.230	19.826
12/16/02	03/17/03	95.020	1.121	12/17/07	03/17/08	93.140	21.444
03/17/03	06/17/03	94.780	1.533	03/17/08	06/17/08	93.180	23.128
06/16/03	09/16/03	94.514	2.005	06/16/08	09/16/08	93.150	24.870
09/15/03	12/15/03	94.315	2.539	09/15/08	12/15/08	93.125	26.665
12/15/03	03/15/04	94.100	3.138	12/15/08	03/16/09	93.035	28.537
03/15/04	06/15/04	94.030	3.803	03/16/09	06/16/09	93.075	30.474
06/14/04	09/14/04	93.905	4.529	06/15/09	09/15/09	93.050	32.468
09/13/04	12/13/04	93.800	5.314	09/14/09	12/14/09	93.030	34.514
12/13/04	03/14/05	93.660	6.166	12/14/09	03/15/10	92.940	36.641

Table A.5
(*Continued*)

Delivery	Maturity	Price	Convex.	Delivery	Maturity	Price	Convex.
03/14/05	06/14/05	93.650	7.085	03/15/10	06/15/10	92.980	38.831
06/13/05	09/13/05	93.575	8.064	06/14/10	09/14/10	92.950	41.079
09/19/05	12/19/05	93.505	9.183	09/13/10	12/13/10	92.930	43.376
12/19/05	03/20/06	93.395	10.294	12/13/10	03/14/11	92.845	45.756
03/13/06	06/13/06	93.420	11.379	03/14/11	06/14/11	92.885	48.199
06/19/06	09/19/06	93.375	12.707	06/13/11	09/13/11	92.855	50.700

Table A.6
51-Instruments portfolio—Eurodollar deposits & swaps

Settle	Maturity	Rate	Settle	Maturity	Rate
08/03/01	09/03/01	0.0366	08/03/01	08/03/02	0.0385
08/03/01	11/03/01	0.0359	08/03/01	08/03/03	0.0444
08/03/01	02/03/02	0.0360	08/03/01	08/03/04	0.0491
			08/03/01	08/03/05	0.0523
			08/03/01	08/03/06	0.0547
			08/03/01	08/03/08	0.0576
			08/03/01	08/03/11	0.0598
			08/03/01	08/03/31	0.0632

References

Adams, K.J., Van Deventer, D.R., 1994. Fitting yield curves and forward rate curves with maximum smoothness. Journal of Fixed Income, 52–62.

Black, F., Derman, E., Toy, W., 1990. A one-factor model of interest rates and its application to treasury bond options. Financial Analysts Journal 1, 33–39.

Brace, A., Gatarek, D., Musiela, M., 1997. The market model of interest rate dynamics. Mathematical Finance 7, 27–155.

BIS, 1999. Zero-coupon yield curves. Technical documentation. Bank for International Settlements, Basel.

Boyd, S., Kim, S., Vandenberghe, L., Hassibi, A., 2004. A tutorial on geometric programming. Technical Report.

Brigo, D., Capitani, C., Mercurio, F. 2001. On the joint calibration of the libor market model to caps and swaptions market volatilities. Working Paper, Banco IMI, San Paolo-IMI Group, pp. 1–21.

Buono, M., Gregoru-Allen, R.B., Yaari, U., 1992. The efficacy of term structure estimation techniques: A Monte Carlo study. Journal of Fixed Income 1, 52–59.

De Jong, F., Driessen, J., Pelsser, A., 2001. Libor market models versus swap market models for pricing interest rate derivatives: An empirical analysis. European Finance Review 5, 201–237.

Deaves, R., Parlar, M., 1999. A generalized bootstrap method to determine the yield curve. Technical Report, DeGroot School of Business, McMaster University.

Dupačová, J., Bertocchi, M., 2001. From data to model and back to data: A bond portfolio management problems. European Journal of Operations Research 134, 261–278.

Dupačová, J., Abaffy, J., Bertocchi, M., Hušková, M., 1997. On estimating the yield and volatility curves. Kybernetika 33, 659–673.

Hildebrand, F.B., 1987. Introduction to Numerical Analysis. Dover Publications Inc., New York.

Hull, J., 2000. Options, Futures, and Other Derivatives, fourth ed. Prentice Hall, New York.

Hull, J., White, A., 2000. Forward rate volatilities, swap rate volatilities, and the implementation of the libor market model. Journal of Fixed Income 10, 46–62.

Jamshidian, F., 1997. Libor and swap market models and measures. Finance and Stochastics 1, 293–330.

Kortanek, K., 2003. Extracting the zeros under a law of motion by geometric programming. Synopsis for a talk at Carnegie Mellon, December 2003.

Kortanek, K., Medvedev, V., 1999. Models for estimating the structure of interest rates from observations of yield curves. In: Avellaneda, M. (Ed.), Quantitative Analysis in Financial Markets. World Scientific, Singapore, pp. 53–120.

Kortanek, K., Medvedev, V., 2001. Building and Using Dynamic Interest Rate Models. Wiley & Sons, New York.

Kortanek, K., Yunianto, H., 2004. Analyzing the Indonesian Government bond market with a new coupon stripping model and principal components. In: 8th International Congreee on Insurance Mathematics & Economics, June 2004, Rome.

Nelson, C., Siegel, A., 1987. Parsimonious modeling of the yield curve. Journal of Business 60, 473–489.

Shea, G.S., 1985. Term structure estimation with exponential splines. Journal of Finance 40, 319–325.

Svensson, L., 1995. Estimating forward interest rates with the extended Nelson & Siegel method. Sveriges Riskbank Quarterly Review 3, 13–26.

Vasicek, O.A., Fong, H.G., 1977. Term structure modeling using exponential splines. Journal of Finance 37, 339–356.

Wets, R., Bianchi, S., Yang, L., 2001. Serious zero-curves. http://epirisk.com.

Chapter 3

PROTECTING INVESTORS AGAINST CHANGES IN INTEREST RATES

OLIVIER DE LA GRANDVILLE*

Department of Economics, University of Geneva, 40 Boulevard du Pont-d'Arve, 1211 Genève 4, Switzerland
e-mail: Olivier.DeLaGrandville@ecopo.unige.ch

Contents

1. Basic concepts for valuation and immunization of bond portfolios in continuous time 73
   1.1. The instantaneous forward rate 73
   1.2. The continuously compounded spot rate 75
   1.3. Introducing the missing link: The continuously compounded total return 77
      1.3.1. Definition of the continuously compounded total return 79
      1.3.2. Property 1 79
      1.3.3. Property 2 79
   1.4. Relationships between the total return, the forward rate and the spot rate 80
   1.5. Theorems on the behavior of the forward rate and the total return 81
   1.6. The spot rate curve as a spline and its corresponding forward rate curve 84
2. Immunization: A first approach 90
   2.1. The continuously compounded horizon rate of return 91
   2.2. A geometrical representation of the horizon rate of return 91
   2.3. Existence and characteristics of an immunizing horizon 93
   2.4. The Macaulay concept of duration, its properties and uses 94
      2.4.1. Definition 94
      2.4.2. Properties of the Macaulay duration 97

* I wish to thank a number of Institutions and persons for the support they have provided to this study: Lombard Odier, and especially Thierry Lombard, Patrick Odier, Ileana Regly and Philippe Sarrasin; the TIAA–CREF Institute, and particularly Douglas Fore. It is with equal pleasure that I express my gratitude to Ken Clements, Jean-Marie Grether, Jochen Kleinknecht, Henri Loubergé, Michael McAleer, Paul Miller, Elizabeth Paté-Cornell, Brigitte van Baalen, Jürg Weber and Milad Zarin-Nejadan. Above all, I wish to thank my colleague of University of Western Australia Anthony Pakes for numerous highly useful suggestions and insightful comments.
This study was partly carried out while I was visiting the Department of Management Science and Engineering at Stanford University and the Department of Economics at the University of Western Australia. Both institutions should be thanked for their hospitality and their stimulating working atmosphere.

Handbook of Asset and Liability Management, Volume 1
Edited by S.A. Zenios and W.T. Ziemba
Copyright © 2006 Elsevier B.V. All rights reserved
DOI: 10.1016/S1872-0978(06)01003-9

2.5. A second-order condition 99
2.6. The immunization problem 100
3. Protecting investors against any shift in the interest rate structure—A general immunization theorem 102
3.1. Notation 102
3.2. Present values at time 0 104
3.3. Future values at time 0 105
3.4. Present values at time ε 105
3.5. Future values at time ε 106
3.6. Further concepts for immunization: the moments of order k of a bond and a bond portfolio 106
3.6.1. The moment of order k of a bond 106
3.6.2. The moment of order k of a bond portfolio 106
3.6.3. Properties of moments 107
3.7. A general immunization theorem 109
Proof and discussion 109
3.8. The nature of the cash flows of an immunizing portfolio 118
4. Applications 118
4.1. The spot structures and their shifts 119
4.2. Building immunizing portfolios 122
4.3. Immunization results 124
4.4. How large should we set the immunization parameter K? 126
4.5. Infinity of solutions 128
4.6. How sensitive are immunizing portfolios to changes in horizon H? 130
4.7. How sensitive are immunizing portfolios to a change in the basket of available bonds? 131
5. Conclusion and suggestions 133
6. Notes to references 137
References 137

Individual and collective security are among the gauges of civilization, and at the same time they constitute necessary conditions for its progress. Their systematic pursuit is relatively recent among nations: prompted by the dreadful lessons of the 20th century, they took their present shape after World War II. Only at that time serious consideration started being given to social and economic security, which resulted into insurance and pension plans being widespread. Large savings, concomitant to decades of economic progress, were invested into default-free vehicles, by those institutions which were looking for steady, fixed, income flows.

In the world of fixed exchange rates which followed the Bretton-Woods agreements, it was only natural that interest rates remained relatively stable. When that state of affairs was disrupted some three decades later, volatile interest rates transformed so-called fixed income assets into risky investments. It is at that time that considerable research was undertaken to protect institutional investors against the fluctuations of interest rates and of bond prices. The word "immunization" describes the steps taken by a bond manager to build up a portfolio which will be minimally sensitive to interest rate shifts. This study has two aims: to give a comprehensive presentation of the immunization problem; to provide novel, original results. In particular, we will show that it is possible to protect investors against any change in the term structure of interest rates, and we will suggest a method to achieve this.

Let us first explain briefly what is at stake. What are the exact consequences of a rise, for instance, in interest rates? For a bond holder, is it good or bad news? The answer to this question depends crucially on the *horizon* of the investor. It is easy to figure out that an investor with a short horizon will suffer from a rise in interest rates because her bond (or bond portfolio) will immediately decrease in value, and this capital loss may not be compensated by the reinvestment of coupons at a higher rate. On the other hand, an investor with a longer horizon may welcome such a rise in interest rates for the two following reasons: first, the initial capital loss will start eroding as bond prices return to their par value, and second, coupons will be reinvested at a higher rate. We may then surmise that, under special circumstances, an intermediate horizon exists for which the investor may be indifferent to a change in interest rates because capital gains (or losses) are nearly exactly compensated by losses (or gains) in coupon reinvestment.

Such a horizon exists, indeed; it is equal to the *duration* of the portfolio's bond, which is the weighted average of the portfolio's times of payment, the weights being the shares of the portfolio's present value cash flows in the portfolio's total value. This prompts to ask the natural question: reversing the problem, consider an investor's horizon as *given*. Is it possible to constitute a bond portfolio such that the investor is protected (immunized) against *any* change in the interest rate structure? The problem was well understood and tackled a few decades ago under simple hypotheses: the interest rate structure (the interest rates corresponding to various maturities) was supposed to receive parallel shifts. The recipe was basically to build a portfolio such that its duration was equal to the investor's horizon. In order to take into account the fact that short rates were usually more volatile than long ones, it was later proposed to equate not only the first moment (the duration) of the portfolio to the horizon, but some of its successive

moments to the successive integer powers of the horizon (see Nawalkha and Chambers (1999b))—a portfolio's moment of order k is the natural generalization of the concept of duration: it is the weighted average of the kth power of its times of payment.

No analytic foundation, however, seems to have been given to the question of determining how many moments should be taken into consideration. Indeed, it seems that no link was established between first- and second-order conditions for immunization, on the one hand, and the number of moments to be considered. It is even difficult to figure out how second-order conditions for a minimum of the portfolio's future value have been brought into the picture, if at all. Finally, present theory is silent on the question of the minimum number of bonds to include in any immunizing portfolio. This study intends to fill in those gaps. To do so, we will propose a general method which will always lead to a local minimum of the portfolio's future value.

Our first task will be to lay out carefully our tools of analysis: the instantaneous forward rate and the continuously compounded spot rate (Section 1). The reason for paying particular attention to those tools is that, in our opinion, they are not well understood, as the following examples illustrate. First consider the well-known, usual, shape of the spot rate curve (the spot rate as a function of maturity): increasing, concave, leveling off to a plateau. It does not seem to be realized that the corresponding forward curve *must be decreasing* before the spot rate reaches its plateau. This failure is evidenced by the various descriptions of the companion forward curve: the fruit of remarkable imagination, they unfortunately never exhibit the right properties. On a more general level—and this may explain the shortcomings just mentioned—it is not recognized that the forward instantaneous rate and the continuously compounded spot rate play *exactly* the same roles as a marginal and an average quantity in economics. The reason for this poor perception may be the following: while marginal and average values in economics are derived from well-known quantities (for example, total cost, or total revenue), the common variable giving rise here to marginal and average values is well-hidden: to the best of our knowledge, it has not yet even been defined. We will need to introduce it, and we will call it the "continuously compounded *total* return".

In Section 2, we present the concept of immunization in the following, simple framework. We will consider that the spot rate curve undergoes parallel shifts. We will then show that under those special circumstances duration plays indeed a central role in the immunization process, in the sense that a portfolio is completely immunized if its duration is equal to the investor's horizon.

Section 3 is devoted to stating and proving a general immunization theorem. In order to cope with real world cases, where the spot rate curve can receive any kind of shift, it is necessary to recall the concept of the kth moment of a bond portfolio. Its properties are essential, and are carefully spelt out. The key to successful immunization is to build a portfolio such that the gradient of its future value is zero, and such that its Hessian matrix is positive definite. To achieve this last aim, we suggest to equate the last element of the Hessian matrix to a positive, arbitrary, parameter. We examine the important consequences of these conditions in terms of the total number of bonds to be considered in the immunizing portfolio, and of the structure of the portfolio. Finally, the theorem

will be put to the test in Section 4, where numerous applications are given. We will consider completely arbitrary initial spot rate curves, as well as large, arbitrary shifts of those.

1. Basic concepts for valuation and immunization of bond portfolios in continuous time

1.1. The instantaneous forward rate

Consider a time span $[u, v]$ subdivided into n intervals $\Delta z_1, \ldots, \Delta z_j, \ldots, \Delta z_n$. This partitioning of $[u, v]$ is arbitrary in the following sense: each Δz_j interval ($j = 1, \ldots, n$) has arbitrary length; two given intervals, say Δz_1 and Δz_2, may or may not be equal. Such an arbitrary partitioning of the time span $[u, v]$ is represented in Figure 1.

Suppose that to each interval Δz_j corresponds a forward rate. Such a forward rate is defined as the *yearly* interest rate for a loan agreed upon at time u; the loan is to start at the beginning of the Δz_j interval, and is to be paid back at the end of Δz_j. Call f_j this forward rate. For instance, f_j may be equal to 5% per year.

If \$1 is lent at the beginning of this time interval Δz_j, what will be the amount to be reimbursed at the end of Δz_j? There are two ways of answering this question. One could agree that it is

- either $(1+f_j)^{\Delta z_j}$, in conformity with annual interest compounding if Δz_j is typically equal to a few years;
- or $1 + f_j \Delta z_j$, following what is usually done if the interval Δz_j is smaller than one year, which will be justified below.

Until now, we have said nothing of the length of the time span $[u, v]$. In fact, inasmuch as the partitioning of $[u, v]$ is arbitrary, so is its length: it could be a non integer number of years or days. It turns out that, whatever this length may be, we will at one point consider that the number n of intervals Δz_j tends towards infinity, and that at the same time the largest of those intervals tends towards zero. Therefore, we can assume already at this stage that Δz_j is small. We will now show why it is both natural and legitimate to use the second formula, i.e., $1 + f_j \Delta z_j$.

We could first observe that, for any fixed value Δz_j, $1 + f_j \Delta z_j$ is a first-order approximation of $(1 + f_j)^{\Delta z_j}$ at point $f_j = 0$. This can be shown by taking the first-order Taylor expansion of $(1 + f_j)^{\Delta z_j}$ around $f_j = 0$. But there is more to it. It turns out that $1 + f_j \Delta z_j$ is an excellent, *second-order*, approximation of $(1 + f_j)^{\Delta z_j}$ around point $(f_j, \Delta z_j) = (0, 0)$. Indeed, consider the latter expression as a function of the two

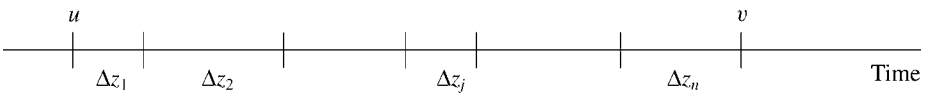


Fig. 1. An arbitrary partitioning of the time span $[u, v]$.

variables f_j and Δz_j, which we will for convenience denote momentarily as f and h respectively. Let $\varphi(.)$ be the function

$$\varphi(f,h) = (1+f)^h \equiv (1+f_j)^{\Delta z_j}. \tag{1}$$

Expand (1) in Taylor series around point $(f,h) = (0,0)$. The second-order approximation is

$$\varphi(f,h) \approx \varphi(0,0) + \frac{\partial\varphi}{\partial f}(0,0)f + \frac{\partial\varphi}{\partial h}(0,0)h + \frac{1}{2}\left[\frac{\partial^2\varphi}{\partial f^2}(0,0)f^2 + 2\frac{\partial^2\varphi}{\partial f\,\partial h}(0,0)fh + \frac{\partial^2\varphi}{\partial h^2}(0,0)h^2\right]. \tag{2}$$

Immediate calculations show that among the six terms in the right-hand side of (2), all are equal to zero except the first one, equal to 1, and the fifth (involving the cross second-order derivative), equal to fh. Therefore, we may write

$$1 + fh \approx (1+f)^h \tag{3}$$

as a *second-order* approximation. From a geometric point of view, $1+fh$ is a surface which has a high (second-order) degree of contact with surface $(1+f)^h$ at point $(0,0)$ because all partial derivatives up to the second-order are equal for each function (as to the tangent plane at $(0,0)$, or the first-order contact surface, it is just $\varphi(f,h) = 1$). To get some sense of this closeness, suppose that f_j is 5% per year and Δz_j is one month, or $0.08\bar{3}$ year. On the one hand, $(1+f_j)^{\Delta z_j} = 1.0041$, while $1 + f_j\Delta z_j = 1.0042$ on the other.

Consider now that the rate of interest f_j is compounded not once within interval Δz_j, but m times. Using this approximation, one dollar invested at the beginning of this interval thus becomes $(1 + f_j\Delta z_j/m)^m$ where $\Delta z_j/m$ is the length of time over which the interest is compounded. Now set $f_j\Delta z_j/m = 1/k$; thus $m = kf_j\Delta z_j$ and one dollar becomes $(1+1/k)^{kf_j\Delta z_j}$. Take the limit of this amount when the number of compoundings m within Δz_j tends towards infinity—equivalently, when $k \to \infty$, since the $f_j\Delta z_j$'s terms are finite numbers. We get

$$\lim_{k\to\infty}\left(1+\frac{1}{k}\right)^{kf_j\Delta z_j} = \mathrm{e}^{f_j\Delta z_j}. \tag{4}$$

This result could have been reached without recourse to the second-order approximation $(1+f_j)^{\Delta z_j} \approx 1 + f_j\Delta z_j$. From formula $(1+f_j)^{\Delta z_j}$, when interest is compounded m times within Δz_j, we get $(1 + f_j/m)^{m\Delta z_j}$ at the end of interval Δz_j. Now set $f_j/m = 1/k$, so that $m = kf_j$. One dollar becomes $(1+1/k)^{kf_j\Delta z_j}$, and we get the above result.

If the same limiting process is taken for each interval Δz_j $(j = 1,\ldots,n)$, an investment C_u at time u becomes

$$C_v = C_u\mathrm{e}^{\sum_{j=1}^{n} f_j\Delta z_j} \tag{5}$$

at time v.

Suppose now that, whatever the initial partitioning of the time span $[u, v]$, the sum $\sum_{j=1}^{n} f_j \Delta z_j$ tends toward a unique limit when the number of time intervals n tends toward infinity, and when the maximum interval length Δz_j tends toward zero. The forward rate function is then said to be integrable, and the limit is the definite integral of f_j over $[u, v]$. We have

$$\lim_{\substack{n\to\infty \\ \max \Delta z_j \to 0}} \sum_{j=1}^{n} f_j \Delta z_j = \int_u^v f(z)\,\mathrm{d}z. \tag{6}$$

We are now ready for a definition of the instantaneous forward rate.

Definition. Let $[u, v]$ be a time span. Let z be any point of time within this interval. The *instantaneous forward rate* $f(z)$ is the yearly interest rate decided upon at time u for a loan starting at time z, $z \in [u, v]$, for an infinitesimally small trading period $\mathrm{d}z$.

Sometimes, to recall the time (u) at which this instantaneous forward rate is agreed upon, the forward rate is denoted $f(u, z)$. If, for instance, the time of inception of the contract is 0 and the time at which the loan starts is T (with an infinitesimal trading period), the instantaneous forward rate is denoted $f(0, T)$, and so forth.

From (5) and (6) a sum C_u invested during $[u, v]$ at the infinite number of instantaneous forward rates $f(u, z)$—or $f(z)$ for short—becomes

$$C_v = C_u \mathrm{e}^{\int_u^v f(z)\,\mathrm{d}z}. \tag{7}$$

A final word about (6) and (7) is in order. Notice the dimension of the infinitesimal element $f(z)\,\mathrm{d}z$ in the integral: since $f(z)$ is in (1/time) units, and since $\mathrm{d}z$ is in time units, $f(z)\,\mathrm{d}z$ is dimensionless; and so are $\int_u^v f(z)\,\mathrm{d}z$ and $\exp[\int_u^v f(z)\,\mathrm{d}z]$.

1.2. The continuously compounded spot rate

Consider the following, particular, rate of interest:

- the time of signing of the contract and the starting point of the loan are at the same point of time: u;
- the length of the loan is the time span $[u, v]$; and
- the amount loaned is C_u; the amount due is C_v.

There are a number of ways of defining the rate of interest on such a loan. If this rate is calculated once only over the whole period $[v - u]$ (which may or may not be an integer number of years), one could define this rate as the *spot* rate (because it corresponds to a loan starting at the very signing of the contract) compounded *once* over $[v - u]$. Denote such a spot rate as $s_{u,v}^{(1)}$. The lower indexes of s are self-explanatory; the superscript (1) refers to the fact that the rate of interest is calculated once over $[u, v]$. We have

$$s_{u,v}^{(1)} = \frac{C_v - C_u}{C_u} / (v - u) \tag{8}$$

and equivalently

$$C_v = C_u\big[1 + s_{u,v}^{(1)}(v-u)\big]. \tag{9}$$

Suppose now that our spot interest is compounded m times over interval $(v-u)$ instead of once only. It implies that the length of time between any two successive compoundings is $(v-u)/m$. The spot rate is now written $s_{u,v}^{(m)}$, and Equation (9) becomes

$$C_v = C_u\big[1 + s_{u,v}^{(m)}(v-u)/m\big]^m. \tag{10}$$

In order for the contracts to be equivalent (that is, to yield the same amount due C_v), the right-hand sides of (9) and (10) must be equal. Hence

$$1 + s_{u,v}^{(1)}(v-u) = \big[1 + s_{u,v}^{(m)}(v-u)/m\big]^m \tag{11}$$

or

$$s_{u,v}^{(m)} = \frac{m}{v-u}\Big\{\big[1 + s_{u,v}^{(1)}(v-u)\big]^{1/m} - 1\Big\}. \tag{12}$$

Replacing m by 1 in the right-hand side of (12) yields $s_{u,v}^{(1)}$ as it should.

Consider now what happens to (10) when the number of compoundings m tends toward infinity. In the same vein as what we did before (Section 1.1), we can replace $s_{u,v}^{(m)}(v-u)/m$ by $1/k$; therefore, with $m = k\, s_{u,v}^{(m)}(v-u)$, we can write

$$C_v = C_u\left(1 + \frac{1}{k}\right)^{k\, s_{u,v}^{(m)}(v-u)}. \tag{13}$$

Taking the limit of C_v when $m \to \infty$ and $k \to \infty$

$$\lim_{\substack{m\to\infty \\ k\to\infty}} C_v = C_u e^{s_{u,v}^{(\infty)}(v-u)}. \tag{14}$$

This leads to the following definition.

Definition. Let $[u, v]$ be a time span. The *continuously compounded spot rate* is the yearly interest rate, denoted $s(u, v)$ that transforms an investment C_u at time u into C_v at time v, the rate being compounded over infinitesimally small time intervals between u and v.

Denote for simplicity $s_{u,v}^{(\infty)} \equiv s(u, v)$. Equation (14) is

$$C_v = C_u e^{s(u,v)(v-u)}. \tag{15}$$

From (15), we deduce

$$s(u, v) = \frac{\ln(C_v/C_u)}{v-u}. \tag{16}$$

There is an all-important relationship between the continuously compounded spot rate $s(u, v)$ (a number) and the forward rate $f(z)$ (a function defined over the interval $[u, v]$). Arbitrage without transaction costs guarantees that the right-hand sides of (7) and (15) must be identical. Thus, from

$$\mathrm{e}^{\int_u^v f(z)\,\mathrm{d}z} = \mathrm{e}^{s(u,v)(v-u)}$$

we get

$$s(u, v) = \frac{\int_u^v f(z)\,\mathrm{d}z}{v - u}. \tag{17}$$

Therefore the spot rate $s(u, v)$ is nothing else than the average value of the forward rate function $f(z)$ over the interval $[u, v]$. In this average the infinitely large number of elements $f(z)$ are all weighted by the infinitely small quantities $\mathrm{d}z/(v - u)$ whose sum is equal to one.

On the other hand, it is always possible to express a forward rate as a function of a spot rate. Write (17) as

$$\int_u^v f(z)\,\mathrm{d}z = (v - u)s(u, v) \tag{18}$$

and take the derivative of (18) with respect to v. This yields

$$f(v) = s(u, v) + (v - u)\frac{\mathrm{d}s(u, v)}{\mathrm{d}v}.$$

The instantaneous forward rate and the continuously compounded spot rate play respectively exactly the roles of a marginal and an average quantity, in economics parlance. This fact does not seem to be known, and for that reason we will now introduce what we feel to be the missing link between the two concepts, namely the continuously compounded *total* return.

1.3. Introducing the missing link: The continuously compounded total return

Let us first simplify our notation. Without loss of generality, call 0 instant u, and T instant v. Our time span $[u, v]$ is now replaced by $[0, T]$. The simplification comes from the fact that T stands both for an *instant* of time and for the *length* of the time span $[0, T]$. With this notation, our spot rate compounded once over period T, corresponding to Equation (8), is now written

$$s_{0,T}^{(1)} = \frac{C_T - C_0}{C_0}\Big/T \tag{19}$$

implying

$$s_{0,T}^{(1)} T = \frac{C_T - C_0}{C_0}.$$

This last quantity can be called the total return compounded once over period T. Contrary to the forward rate and the spot rate, which were expressed in (1/time) units, this magnitude is unitless. (To emphasize this fact, we have not called $s^{(1)}_{0,T}T$ the total *rate* of return, but the total return.) Equation (19) implies

$$C_T = C_0\big[1 + s^{(1)}_{0,T}T\big]. \tag{20}$$

Suppose now that this total return corresponds to a spot rate compounded n times over the same interval T. It is thus defined as the product of that spot rate and the time length T. Call $s^{(n)}_{0,T}$ the corresponding rate of return per year. The length of time after which this rate is compounded is T/n. Therefore Equation (20) is transformed into

$$C_T = C_0\left(1 + s^{(n)}_{0,T}\frac{T}{n}\right)^n. \tag{21}$$

The total return when the spot rate is compounded n times over T is denoted as $s^{(n)}_{0,T}T$, and remains *dimensionless*.

From (21) and (20), we have

$$s^{(n)}_{0,T}T = n\big[(C_T/C_0)^{1/n} - 1\big] = n\big[(1 + s^{(1)}_{0,T}T)^{1/n} - 1\big]. \tag{22}$$

Setting $n = 1$ into (22), we verify that we get (19) as we should.

Let $n \to \infty$. We will now determine the limit of the corresponding, unitless, total return $s^{(\infty)}_{0,T}T$. We have to calculate

$$s^{(\infty)}_{0,T}T = \lim_{n\to\infty} s^{(n)}_{0,T}T = \lim_{n\to\infty} n\big[(C_T/C_0)^{1/n} - 1\big] \tag{23}$$

which will be called the *continuously compounded total return*.

Set $n = 1/\alpha$; then

$$s^{(\infty)}_{0,T}T = \lim_{n\to\infty} s^{(n)}_{0,T}T = \lim_{\alpha\to 0}\big\{\big[(C_T/C_0)^{\alpha} - 1\big]/\alpha\big\}. \tag{24}$$

Applying L'Hospital's rule to (24) yields

$$s^{(\infty)}_{0,T}T = \lim_{\alpha\to 0}\big[(C_T/C_0)^{\alpha}\ln(C_T/C_0)\big] = \ln(C_T/C_0). \tag{25}$$

Since $s^{(\infty)}_{0,T} \equiv s(0,T)$, the continuously compounded total return $s(0,T)T$ will be denoted $S(0,T)$ for short. Thus

$$S(0,T) = \ln(C_T/C_0).$$

Let now $f(0,z)$ be the forward rate decided upon at time 0, for an infinitesimal trading period starting at time z, with z between 0 and T. In the case of an infinite number of forward markets and in the absence of transaction costs, we know that arbitrage will enforce the equalities:

$$C_T = C_0 e^{s(0,T)T} = C_0 e^{\int_0^T f(0,z)\,dz}$$

from which $S(0, T) = \log(C_T/C_0) = s(0, T)T = \int_0^T f(0, z)\,\mathrm{d}z$ results. Notice that the continuously compounded total return is dimensionless, as it should.

We may now state a formal definition of the continuously compounded total return.

1.3.1. Definition of the continuously compounded total return

Let $f(0, z)$ be an instantaneous forward rate agreed upon at 0 for a loan starting at z $(0 \leqslant z < T)$, for an infinitesimal trading period. Define the *continuously compounded total return* at time T as the integral sum of the forward rates between 0 and T, equal to $\int_0^T f(0, z)\,\mathrm{d}z$. It is denoted $S(0, T)$, and is a pure (dimensionless) number. The continuously compounded total return will be abbreviated as the "total return", the continuously compounded spot rate $s(0, T)$ will be abbreviated to "spot rate", and "forward rate" represents the instantaneous forward rate $f(0, T)$.

From what precedes, we can state the following properties:

1.3.2. Property 1

Let $s(0, T)$ be a continuously compounded spot rate for a loan started at time 0 and maturing at T. Then the total return $S(0, T)$ is equal to $s(0, T)T$.

1.3.3. Property 2

Let C_0 denote the value of an investment at time 0; C_T is its value at time T. Then the total return $S(0, T)$ equals $\ln(C_T/C_0)$.

An example may be useful. Suppose the total return is 32% or 0.32 after four years. It means that one dollar becomes $\mathrm{e}^{0.32} = 1.377$ dollars after four years. Equivalently, it also means that

- the continuously compounded yearly return, or the spot rate with 4-years maturity is $0.32/4 = 0.08$ per year;
- the forward rate function for maturities between $t = 0$ and $t = 4$ is *any* function $f(z)$ such that

$$\int_0^4 f(z)\,\mathrm{d}z = 0.32,$$

  or that its average value is $\int_0^4 f(z)\,\mathrm{d}z/4 = 0.08$ per year. In order to get such a function $f(0, z)$, choose *any* differentiable function $s(0, z)$ with $z \in [0, 4]$ that goes through point $(4, 0.08)$. Then $f(0, z)$ is simply $s(0, z) + z\,\mathrm{d}s(0, z)/\mathrm{d}z$;
- the yearly, once-a-year compounded spot rate, $s_{0,4}$, is such that

$$(1 + s_{0,4})^4 = \mathrm{e}^{0.32}$$

  or

$$s_{0,4} = \mathrm{e}^{0.08} - 1 = 0.083 \text{ per year.}$$

Those definitions of the forward rate, the spot rate and the total return enable us to write the following equalities:

$$C_T = C_0 \mathrm{e}^{\int_0^T f(z)\,\mathrm{d}z} = C_0 \mathrm{e}^{s(0,T)T} = C_0 \mathrm{e}^{S(0,T)}. \tag{26}$$

Noticing that the first equality in (26) can be written as

$$C_T = C_0 \mathrm{e}^{[\frac{1}{T}\int_0^T f(z)\,\mathrm{d}z]T},$$

equalities (26) allow for a nice interpretation of the Euler number e. The Euler number, e, is equal to what becomes of \$1 when, equivalently:

- the *average* of the forward rates over period T is $1/T$;
- the spot rate $s(0, T)$ is equal to $1/T$;
- the continuously compounded total return is 1 (100%).

Suppose $T = 25$ years. Then $1/T$ is 4% per year. One dollar becomes $\mathrm{e} = 2.718\ldots$ dollars if the *average* of the forward rates over 25 years—equivalently the spot rate $s(0, 25)$—is 4%. This average of 4% allows for *negative* values of the forward rate inasmuch as the forward rate is a *real* forward rate, defined as the difference between the nominal forward rate and the inflation rate.

We now examine the precise relationships existing between the three concepts. Understanding these relationships seems essential to avoid innumerable erroneous depictions of forward and spot rate curves such as those appearing regularly in otherwise excellent texts on finance. Indeed, all too often the traditional increasing, concave spot curve leveling off to a plateau is accompanied by a forward curve which has simply no relevance to it. The latter's position can be way off the mark, being faulty on many grounds. In particular, before the spot rate reaches a plateau, it is depicted as still increasing while it must be *decreasing*—this is a property that will be demonstrated in Section 1.5. Texts are sometimes graced with hump-shaped spot rate curves; not infrequently the decreasing part of the hump exhibits a subtangent (measured on the ordinate) *larger* than the spot rate, thus entailing nonsensical *negative* nominal forward rates. The subtangent of $s(0, T)$ is defined as the absolute value of the differential $T\,\mathrm{d}s(0, T)/\mathrm{d}T$. If the spot rate is decreasing, we must still have, with $\mathrm{d}s(0, T)/\mathrm{d}T < 0$:

$$f(T) = s(0, T) + T\frac{\mathrm{d}s(0, T)}{\mathrm{d}T} > 0$$

or

$$-T\frac{\mathrm{d}s(0, T)}{\mathrm{d}T} = \left|T\frac{\mathrm{d}s(0, T)}{\mathrm{d}T}\right| < s(0, T).$$

1.4. Relationships between the total return, the forward rate and the spot rate

It is now straightforward to show that the forward rate and the spot rate are, respectively, the marginal and average quantities of a common value, the total return.

First, consider the total return as expressed as an integral sum, and take its derivative with respect to T. We get the forward rate

$$\frac{\mathrm{d}S(0,T)}{\mathrm{d}T} = \frac{\mathrm{d}}{\mathrm{d}T}\left[\int_0^T f(0,z)\,\mathrm{d}z\right] = f(0,T). \tag{27}$$

In economic parlance, the forward rate is simply the marginal return.

Second, take the average value of $S(0,T)$. We get the spot rate

$$\frac{S(0,T)}{T} = \frac{\int_0^T f(0,z)\,\mathrm{d}z}{T} = \frac{s(0,T)T}{T} = s(0,T). \tag{28}$$

Thus the spot rate is the average return.

Taking the derivative of $S(0,T) = s(0,T)T$ with respect to T and using (27) yields

$$f(0,T) = s(0,T) + \frac{\mathrm{d}s(0,T)}{\mathrm{d}T}T. \tag{29}$$

This shows that the forward rate can be constructed geometrically by adding to the spot rate $s(0,T)$ the differential $[\mathrm{d}s(0,T)/\mathrm{d}T]T$ (see Section 1.6).

To avoid some of the errors we alluded to at the end of Section 1.3, it is important to understand the full meaning of the differential $[\mathrm{d}s(0,T)/\mathrm{d}T]T$ of the spot curve. If $s(0,T)$, the average value of the forward rates, is an increasing function of T, the forward rate (the marginal value of the *total return*) is equal to the average value (the spot rate) plus the increase in the average that now applies to *all* infinitesimal time increments from 0 to T. This increase is $[\mathrm{d}s(0,T)/\mathrm{d}T]T$; indeed it is the rate of increase per additional time to maturity $[\mathrm{d}s(0,T)/\mathrm{d}T]$ multiplied by the sum of all increases in maturity from 0 to T, i.e., $\int_0^T \mathrm{d}z = T$. Thus, the geometrical construction of the forward rate from the spot rate as $f(0,T) = s(0,T) + [\mathrm{d}s(0,T)/\mathrm{d}T]T$ makes sound economic sense. As an application, we can understand immediately what is at stake when representing a spot rate curve with a hump: the decreasing part of the hump may be too steep, in the sense that $[\mathrm{d}s(0,T)/\mathrm{d}T]T < -s(0,T)$ over some interval of T, thus entailing $s(0,T)+[\mathrm{d}s(0,T)/\mathrm{d}T]T < 0$ and hence non-sensical negative (nominal) forward rates $f(0,T)$ over that interval.

Figure 2 summarizes the relationships between the three concepts of total return, spot rate and forward rate. There is a one-to-one correspondence between these three concepts, and always two ways of going from one of these to any of the other two: a direct one and an indirect one. These relationships are used in the next two sections.

1.5. Theorems on the behavior of the forward rate and the total return

We will now state and demonstrate a property of the forward rate and the total return that pertains to the most commonly observed behavior of the spot rate.

Theorem 1. *Let the spot rate $s(0,T)$ be represented by an increasing, concave function of maturity T, leveling off at $s(0,\overline{T}) = \bar{s}$ for an abscissa $\overline{T}$, and equal to level $\bar{s}$ beyond $\overline{T}$. Then*:

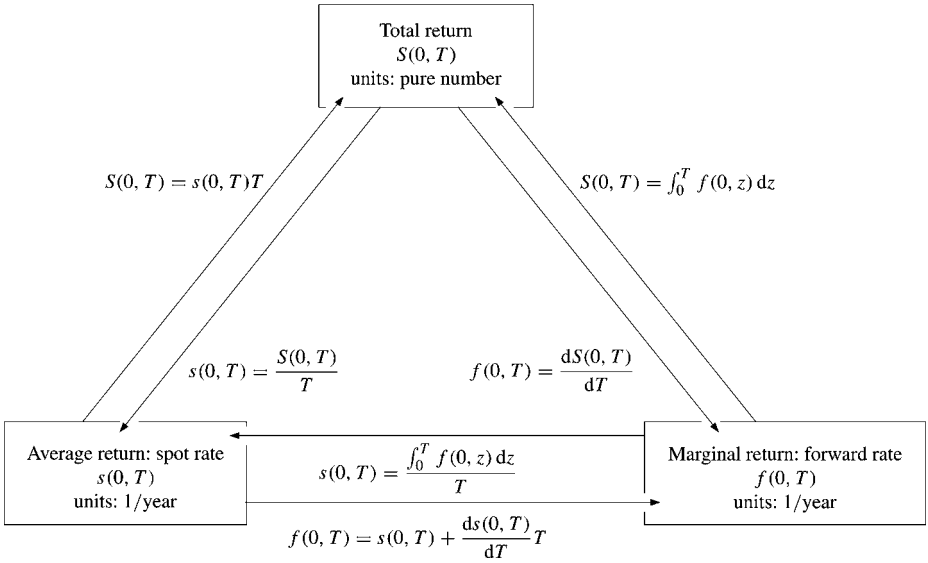


Fig. 2. Summary of the relationships between the total return $S(0, T)$, the spot rate $s(0, T)$ and the forward rate $f(0, T)$.

(a) *the forward rate goes at least through one maximum in the interval* $(0, \overline{T})$ *and is decreasing in an interval to the left of* $\overline{T}$. *If* $T \geqslant \overline{T}$, *the forward rate is equal to the spot rate* $\bar{s}$;
(b) *the total return is an increasing function of* T, *with at least one inflection point for* $0 < T < \overline{T}$. *If* $T \geqslant \overline{T}$, *the total return becomes a linear function of* T.

Proof. Let us first consider the behavior of the slope of the forward rate curve at the origin, i.e., when $T \to 0$. The above assumptions imply $\mathrm{d}s(0, 0)/\mathrm{d}T > 0$. On the other hand, taking the derivative of (34) with respect to T, we deduce

$$\begin{aligned}\frac{\mathrm{d}f(0, T)}{\mathrm{d}T} &= \frac{\mathrm{d}s(0, T)}{\mathrm{d}T} + \frac{\mathrm{d}^2 s(0, T)}{\mathrm{d}T^2}T + \frac{\mathrm{d}s(0, T)}{\mathrm{d}T} \\ &= 2\frac{\mathrm{d}s(0, T)}{\mathrm{d}T} + \frac{\mathrm{d}^2 s(0, T)}{\mathrm{d}T^2}T \end{aligned} \tag{30}$$

and therefore, with $T = 0$

$$\frac{\mathrm{d}f(0, 0)}{\mathrm{d}T} = 2\frac{\mathrm{d}s(0, 0)}{\mathrm{d}T}. \tag{31}$$

Notice how surprisingly simple this intermediate result is: *whatever* the spot rate curve, and in particular *whatever the sign* of the spot curve slope at the origin, the initial slope of the forward rate curve will *always* be twice that of the spot curve.

This result can be generalized while retaining its simplicity. At the origin the nth derivative of the forward rate is $(n+1)$ times the nth derivative of the spot rate, i.e., $\mathrm{d}^n f(0,0)/\mathrm{d}T^n = (n+1)\,\mathrm{d}^n s(0,0)/\mathrm{d}T^n$.

Coming back to (30), we know that for $T = \overline{T}$, $\mathrm{d}s(0,\overline{T})/\mathrm{d}T = 0$ and its left-hand side second derivative is negative ($\mathrm{d}^2 s(0,\overline{T}_-)/\mathrm{d}T^2 < 0$), as implied in our concavity hypothesis. This leads to a left-hand side derivative of $f(0,\overline{T})$ equal to

$$\frac{\mathrm{d}f(0,\overline{T}_-)}{\mathrm{d}T} = \frac{\mathrm{d}^2 s(0,\overline{T}_-)}{\mathrm{d}T^2}\overline{T} < 0$$

and therefore the forward rate is *decreasing* in an interval on the left-hand side of $\overline{T}$.

We know that $\mathrm{d}f(0,0)/\mathrm{d}T$ is strictly positive and that $\mathrm{d}f(0,\overline{T}_-)/\mathrm{d}T$ is strictly negative. It follows that $\mathrm{d}f(0,T)/\mathrm{d}T$ changes sign from positive to negative at least once within the interval $[0,T]$, and therefore that $f(0,T)$ goes through a local maximum at least for one value of the maturity. We denote such a value $\widehat{T}$ ($\widehat{T} \in (0,\overline{T})$)—assuming it is unique.

The final part of part (a) of the theorem is evident from (29): if $\mathrm{d}s(0,T)/\mathrm{d}T = 0$ for $T > \overline{T}$, then $f(0,T) = s(0,T)$.

Part (b) is just a direct consequence of (a). If $f(0,T)$ goes through a unique maximum in the interval $[0,T]$ (at $T = \widehat{T}$), its integral (the total return $S(0,T)$) has an inflection point at $\widehat{T}$; it is a convex function between 0 and $\widehat{T}$ and a concave function between $\widehat{T}$ and $\overline{T}$. Finally, if for $T \geqslant \overline{T}$ $f(0,T)$ is equal to a constant $s(0,T) \equiv \bar{s}$, then

$$\begin{aligned} S(0,T)|_{T\geqslant\overline{T}} &= \int_0^T f(0,z)\,\mathrm{d}z\Big|_{T\geqslant\overline{T}} \\ &= \int_0^{\overline{T}} f(0,z)\,\mathrm{d}z + \int_{\overline{T}}^T f(0,z)\,\mathrm{d}z\Big|_{T\geqslant\overline{T}} \\ &= \bar{s}\overline{T} + \bar{s}T - \bar{s}\overline{T}|_{T\geqslant\overline{T}} = \bar{s}T|_{T\geqslant\overline{T}}, \end{aligned}$$

a *linear* function of maturity T. The proof is thus complete. □

This theorem allows for the following interpretation. Consider the continuous maturity as a discrete variable, albeit with infinitesimally short increments. First, let us explain the property that at the origin the forward rate is equal to the spot rate, and that the forward curve has twice the slope of the spot curve. Let ε ($\varepsilon \to 0$) be the first maturity, 2ε the second one. For only one maturity, the spot rate is the average of one forward rate only; therefore it is equal to it. So $s(0,1) = f(0,0)$. Suppose now that the spot rate for maturity 2ε has increased by an amount Δs. If an average increases by Δs for one additional element, it implies that the second element has increased by twice the increase of the average, i.e., by $2\Delta s$. (The same argument carries, of course, to the case where the average *decreases* by Δs.) Therefore at the origin the slope of the forward rate must be twice the slope of the spot rate. Furthermore, at the point where the spot rate (the average of the forward rates) reaches a maximum, the last forward rate must

equal the spot rate. The fact that between the origin and that point the forward rate goes at least through one maximum is just a consequence of Rolle's theorem applied to the difference between the forward rate and the spot rate.

The property that the forward rate must be decreasing just before the spot rate has reached its maximum (or its plateau) can now be demonstrated as follows. Consider an increasing spot rate. It means that the last forward rate is *above* it: this is what has made the spot rate increase. If the spot rate, for the next maturity, does *not* change (because it has reached a maximum, or a plateau), the forward rate *must* be decreasing. Indeed, suppose the contrary: if the forward rate increases *or* stays constant, the spot rate will still increase, entailing a contradiction with our hypothesis. Therefore, the forward rate can only decrease. Theorem 2 (to follow in the text) has a similar interpretation.

This theorem could be generalized in a natural way to the relationship between the derivative of a function and the average value of that function.

A corollary to this theorem is the following:

Theorem 2. *Let the spot rate be represented by a decreasing, convex function of maturity T, leveling off at $s(0, \overline{T}) = \bar{s}$ for an abscissa $\overline{T}$, and equal to level $\bar{s}$ beyond $\overline{T}$. Then*:

(a) *the forward rate goes at least through one minimum in the interval $(0, T)$ and is increasing in an interval to the left of $\overline{T}$. If $T \geqslant \overline{T}$, the forward rate is equal to the spot rate $\bar{s}$*;

(b) *the total return is an increasing function of T, with at least one inflection point for $0 < T < \overline{T}$. If $T \geqslant \overline{T}$, the total return becomes a linear function of T.*

Proof. The proof follows the same lines as above. □

1.6. The spot rate curve as a spline and its corresponding forward rate curve

In order to derive a correct forward curve corresponding to a spot rate curve, we could proceed geometrically and obtain a fair result by adding to the spot rate its differential $T \cdot \mathrm{d}s(0, T)/\mathrm{d}T$. Consider any value of the maturity, for instance T_0. For this value of the abscissa, the differential $\frac{\mathrm{d}s}{\mathrm{d}T}(0, T_0)T_0$ is equal to distance δ drawn on the ordinate (Figure 3). Adding δ to the spot rate $s(0, T_0)$ yields the forward rate $f(0, T_0)$.

Conversely, from any given forward curve $f(0, T)$ the spot rate can be sketched in the following way. From point T_0 on the abscissa draw the vertical with ordinate $f(0, T_0)$. Consider a series of horizontal lines of length T_0 between height $f(0, T_0)$ and the abscissa. Each of those lines delineates two areas: one below the forward curve and above the horizontal; the other one below the horizontal and above the forward curve. The height of the horizontal for which both areas (A_2 and A_1 in our diagram) are equal yields the spot rate $s(0, T_0)$ (you have just applied the property $s(0, T_0)T_0 = \int_0^{T_0} f(0, z)\,\mathrm{d}z$). Furthermore, not only do you get a fair idea of the spot rate, but of the slope of the spot curve as well at that point. Indeed, consider the difference δ between the forward curve

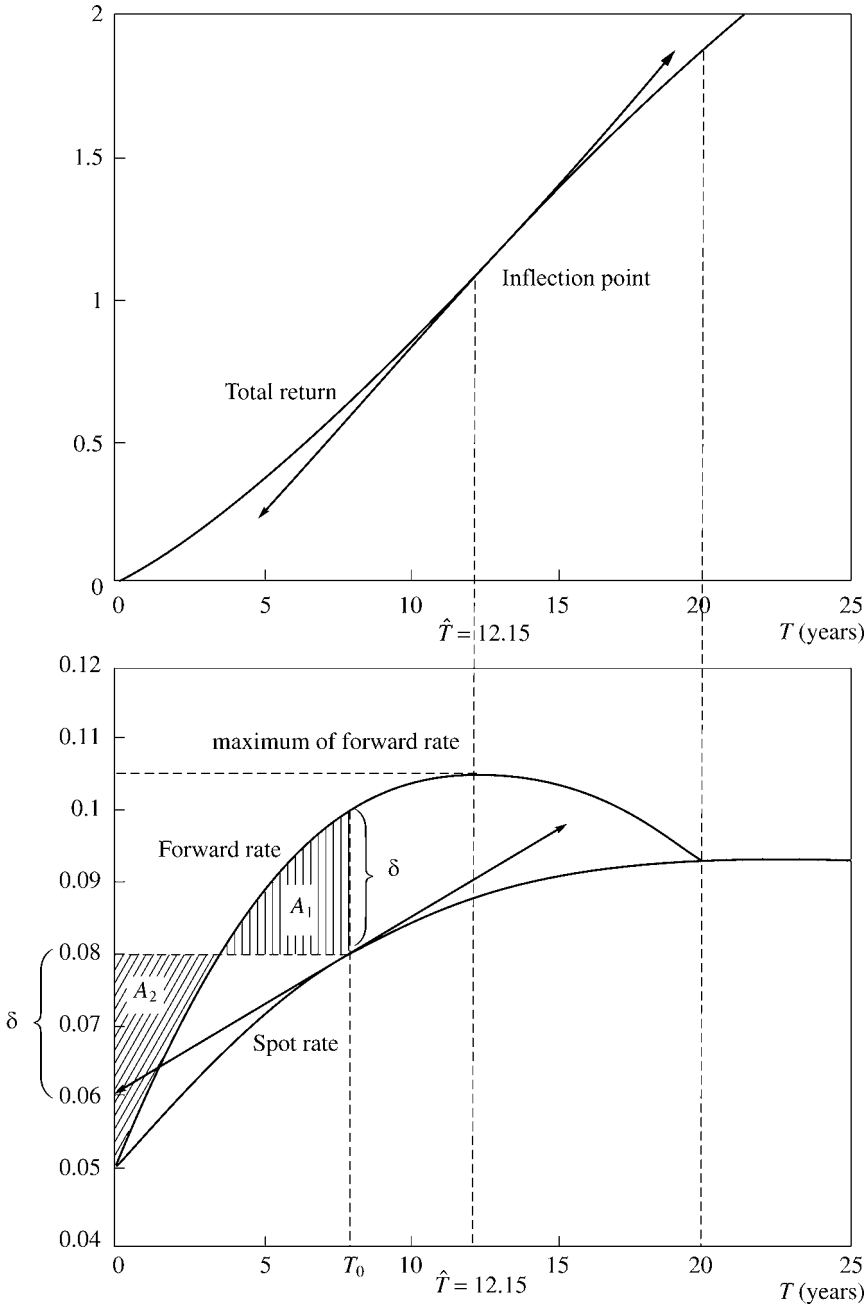


Fig. 3. The continuously compounded total return rate, and the instantaneous forward rate.

and the spot curve at T_0; subtract δ from $s(0, T_0)$ on the ordinate. The line between that point and $(T_0, s(0, T_0))$ is the tangent of the spot curve at T_0.

But it would be more rewarding to verify with precision the property we developed under Theorem 1. To that effect, we will proceed analytically and consider the spot rate curve as a spline represented by a third order polynomial starting at point (0, 0.05) and ending at point (20, 0.0929), with a slope equal to zero at that end point. (On splines, see de Boor (1978) and Wahba (1990).)

The reason for which practitioners use third order polynomials for splines is little known, and merits mentioning here. It stems from the following important property. If a polynomial represents a lath joining a given set of points, a third-order polynomial minimizes both the curvature of the lath and its deformation energy. Indeed, either of these concepts is basically represented—within the confines of affine transformations—by the functional

$$I\big[s(z)\big] = \int_a^b \big[s''(z)\big]^2 \,\mathrm{d}z \tag{32}$$

where $s''(z)$ is the second derivative of the spot rate—or of the spline (denoted $s(z)$ for convenience).

Minimizing the integral I is a particular case of minimizing a functional

$$J\big[s(z)\big] = \int_a^b F\big[z, s(z), s'(z), s''(z)\big]\,\mathrm{d}x.$$

Such a problem is solved by the Euler–Poisson 4th order differential equation

$$\frac{\partial F}{\partial s}(z, s, s', s'') - \frac{\mathrm{d}}{\mathrm{d}z}\frac{\partial F}{\partial s'}(z, s, s', s'') + \frac{\mathrm{d}^2}{\mathrm{d}z^2}\frac{\partial F}{\partial s''}(z, s, s', s'') = 0. \tag{33}$$

In our case (Equation (32)) the arguments z, s and s' are missing in F, which depends solely upon s''. So the Euler–Poisson equation (33) becomes

$$\frac{\mathrm{d}^2}{\mathrm{d}z^2}\frac{\partial F}{\partial s''}(s'') = 0. \tag{34}$$

Applying (34) to (32), yields

$$\frac{\partial F}{\partial s''}(s'') = 2s''(z) \tag{35}$$

and therefore a first-order condition for $s(z)$ to minimize the functional I is that it solves the 4th order differential equation

$$\frac{\mathrm{d}^2}{\mathrm{d}z^2}\big[2s''(z)\big] = 2s^{(4)}(z) = 0, \tag{36}$$

which leads to the third-order polynomial

$$s(z) = \alpha_3 z^3 + \alpha_2 z^2 + \alpha_1 z_1 + \alpha_0. \tag{37}$$

We will now adjust this polynomial to the curve depicted in Figure 3. To that effect, let us revert to our notation

$$s(0, T) = \alpha_3 T^3 + \alpha_2 T^2 + \alpha_1 T + \alpha_0. \tag{37a}$$

Since the spot rate curve is supposed to level off at its end point, only three points will be needed to adjust the curve: indeed, together with the first condition this will imply four equations in the four unknowns a_0, a_1, a_2 and a_3. The first three equations are determined by the knowledge of the initial point, any intermediate point and the end point. Let these three points be denoted as

$$\begin{aligned}
&\big(T_0, s(0, T_0)\big) \equiv (T_0, s_0),\\
&\big(T_1, s(0, T_1)\big) \equiv (T_1, s_1),\\
&\big(\overline{T}, s(0, \overline{T})\big) \equiv (\overline{T}, \bar{s}).
\end{aligned}$$

Constraining the spot curve to go through these three points leads to the first three equations of system (38) below; the last equation of (38) reflects the fact that the derivative of the spot curve at $T = \overline{T}$ is equal to zero. The system of four equations in the four unknowns $\alpha_0, \alpha_1, \alpha_2, \alpha_3$ is

$$\begin{aligned}
&\alpha_3 T_0^3 + \alpha_2 T_0^2 + \alpha_1 T_0 + \alpha_0 = s_0,\\
&\alpha_3 T_1^3 + \alpha_2 T_1^2 + \alpha_1 T_1 + \alpha_0 = s_1,\\
&\alpha_3 \overline{T}^3 + \alpha_2 \overline{T}^2 + \alpha_1 \overline{T} + \alpha_0 = \bar{s},\\
&3\alpha_3 \overline{T}^2 + 2\alpha_2 \overline{T} + \alpha_1 = 0.
\end{aligned} \tag{38}$$

Using the simplifying notation

$$\begin{aligned}
&\boldsymbol{\alpha} = (\alpha_3, \alpha_2, \alpha_1, \alpha_0)',\\
&\mathbf{s} = (s_0, s_1, \bar{s}, 0)',\\
&\mathbf{T} = \begin{bmatrix} T_0^3 & T_0^2 & T_0 & 1\\ T_1^3 & T_1^2 & T_1 & 1\\ \overline{T}^3 & \overline{T}^2 & \overline{T} & 1\\ 3\overline{T}^2 & 2\overline{T} & 1 & 0 \end{bmatrix}
\end{aligned}$$

the vector of unknowns $\boldsymbol{\alpha}$ of the polynomial is such that

$$\mathbf{T} \cdot \boldsymbol{\alpha} = \mathbf{s} \tag{39}$$

and therefore $\boldsymbol{\alpha}$ is equal to

$$\boldsymbol{\alpha} = \mathbf{T}^{-1} \cdot \mathbf{s} \tag{40}$$

provided $\mathbf{T}$ is not singular.

In the example of Figure 3, the three points are

$$(T_0, s_0) = (0, 0.05), \quad (T_1, s_1) = (5, 0.0712), \quad (\overline{T}, \bar{s}) = (20, 0.0929),$$

(these are rounded values. The actual ordinates s_1 and $\bar{s}_2$ used in the calculations were 0.071241 and 0.092857. Throughout this study all results are rounded, but the actual calculations are carried out with 20 decimals). We are then led to the spot rate curve

$$s(0,T) = \begin{cases} 2.21(10^{-6})T^3 - 0.000196T^2 + 0.005171T + 0.05 & \text{if } T \leqslant \overline{T}, \\ s(0,T) = 0.0929 & \text{if } T > 20, \end{cases} \tag{41}$$

from which the following forward rate results:

$$f(0,T) = \begin{cases} 8.86(10^{-6})T^3 - 0.000587T^2 + 0.01034T, \\ f(0,T) = 0.0929 \quad (T > 20). \end{cases} \tag{42}$$

It is immediate to show that $f(0,T)$ has indeed the kind of behavior that was forecast. It goes through a *maximum* at an abscissa $\widehat{T} = 12.146$; from $\widehat{T}$ to $\overline{T}$ it is a *decreasing* function of maturity.

Figure 3 illustrates the behavior of the forward curve corresponding to the spot curve, together with the total return curve.

In accordance with (41) and (42), the total return is

$$\begin{aligned} S(0,T) = s(0,T)\cdot T &= \int_0^T f(0,z)\,\mathrm{d}z \\ &= \begin{cases} 2.21(10^{-6})T^4 - 0.000196T^3 + 0.005171T^2 + 0.05T \\ \qquad\qquad (\text{if } 0 \leqslant T \leqslant 20), \\ 0.0928T \quad (\text{if } T \geqslant 20). \end{cases} \end{aligned} \tag{43}$$

Continuously increasing, with a continuous first derivative, the total return has an inflection point at $\widehat{T} = 12.146$.

A word of caution is in order here: we have shown that the forward rate curve went at least through one maximum for $T < \overline{T}$, but we have said nothing about its concavity or convexity. In the example of Figure 3, the forward curve implied by the concave spot curve turned out to be concave throughout, but this is not at all a general rule. Local convexity can be the case if the following occurs. Consider the derivative of (30). We get

$$\frac{\mathrm{d}^2 f(0,T)}{\mathrm{d}T^2} = 3\frac{\mathrm{d}^2 s(0,T)}{\mathrm{d}T^2} + \frac{\mathrm{d}^3 s(0,T)}{\mathrm{d}T^3}T. \tag{44}$$

At the origin ($T = 0$) a concave spot curve ($\mathrm{d}^2 s(0,0)/\mathrm{d}T^2 < 0$) will always entail a concave forward curve ($\mathrm{d}^2 f(0,0)/\mathrm{d}T^2 < 0$). But for $T > 0$, the forward curve will be convex ($\mathrm{d}^2 f(0,T)/\mathrm{d}T^2 > 0$) if and only if

$$\frac{\mathrm{d}^3 s(0,T)}{\mathrm{d}T^3} > -\frac{3}{T}\frac{\mathrm{d}^2 s(0,T)}{\mathrm{d}T^2} \tag{45}$$

which implies, since $\mathrm{d}^2 s(0,T)/\mathrm{d}T^2 < 0$, $\frac{\mathrm{d}^3 s(0,T)}{\mathrm{d}T^3} / \frac{\mathrm{d}^2 s(0,T)}{\mathrm{d}T^2} < -\frac{3}{T}$.

Denote

$$\frac{\mathrm{d}^3 s(0,T)}{\mathrm{d}T^3} \Big/ \frac{\mathrm{d}^2 s(0,T)}{\mathrm{d}T^2} = s'''/s'' = \frac{\mathrm{d}\ln s''}{\mathrm{d}T} \equiv c$$

the *relative rate of increase* in concavity of $s(0,T)$. Condition (45) amounts to $c < -3/T$.

It is quite interesting to observe how sensitive this condition is to a very small variation in the spot rate. We will suppose that one point only of our original spline is modified by changing point (5, 0.0712) into (5, 0.0747). The spline, always concave, becomes

$$s(0,T) = 5.3\big(10^{-6}\big)T^3 - 0.00032T^2 + 0.0064T + 0.05 \tag{46}$$

and the forward curve becomes

$$f(0,T) = 0.000021T^3 - 0.00096T^2 + 0.01282T + 0.05. \tag{47}$$

This time the forward rate goes through an inflection point at $T^* = 15.045$ years, and becomes convex between T^* and T, as Figure 4 shows. (This corresponds to the solution of equation $c = s'''(0,T)/s''(0,T) = -3/T$.) Only precise calculations of condition (45) lead to this conclusion, which cannot be obtained by simple inspection of the spot rate curve.

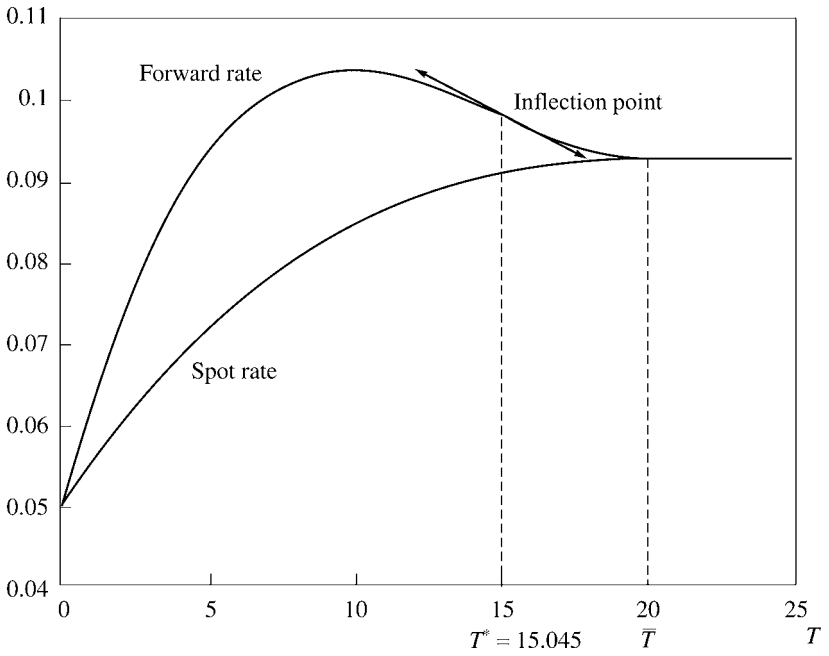


Fig. 4. A concave spot rate curve does not necessarily entail a concave forward rate curve.

We close this section by drawing attention upon a rather surprising consequence of Theorem 1. The usual, innocuous looking, increasing and concave spot rate curve covers a well-hidden, important message regarding the efficiency of investments as a function of their *horizon*. Indeed, it implies a total return on investments which has the exact outlook of a one-variable S-shaped production function with an inflection point, where the variable is neither capital nor labor, but the investment's horizon (see Theorem 1 in Section 1.5, illustrated in Figure 3). For short and medium term projects, this marginal efficiency is first increasing; it is then decreasing. For very long horizons the horizontal spot rate curve, equal to the forward rate, makes the case of constant returns.

These observations prompt us to indulge in the following conjecture. An increasing marginal efficiency of short and medium term investments followed by decreasing marginal efficiency of longer investments may help track down the ever-elusive causes of the "normal" shape of the interest rate term structure. It is my belief that theories of the interest term structure have been somewhat lopsided until now, centering essentially on the behavior of agents in the financial markets, thus giving short shrift to its real counterpart. A natural question to ask now is whether the increasing and decreasing marginal efficiency of investments according to their horizon stands the trial of facts.

2. Immunization: A first approach

In the introduction of this chapter, we have shown the essential role of the investor's horizon when evaluating the consequences of any change in the rates of interest. We had in mind two objects of interest: first, the instantaneous change in the bond portfolio's value due to the drop or the rise of interest rates; and second, the relative decrease or increase, respectively, of the reinvestment proceeds of the coupons. We can now be more specific, and define the future value, at horizon H, of the portfolio: this will take into account both effects at the same time.

However, this measure of the future bond performance can still be improved upon the following reason. The attractiveness of an investment in terms of its future value is not easy to determine. For instance, an investment's future value equal to 136, in $H = 6$ years, when the present value is 94, is difficult to interpret. It is much more convenient to read the investment's performance in terms of a yearly rate of return, either compounded once a year, or continuously compounded. If compounded once a year, the yearly return is $(136/94)^{1/6} - 1 = 6.35\%$; continuously compounded, the yearly return is $(1/6)\ln(136/94) = 6.16\%$. We now introduce in a precise way the horizon rate of return. We have at least two choices: we can consider it either compounded once a year, or continuously compounded. Continuously compounded returns are much more convenient to deal with than any other form of return, so we use those.

To have a clear idea of immunization, suppose that the spot rate structure is flat, and that it undergoes a parallel shift. We can use the tools we are now familiar with, and set

in the spot rate polynomial

$$s(0, T) = a_1 + a_2 T + a_3 T^2 + a_4 T^3, \tag{48}$$

$a_2 = a_3 = a_4 = 0$, retaining only a_1. So $s(0, T) = a_1$ at time 0 when a bond or a bond portfolio is bought. A very short time later, at time ε, the spot structure shifts to a new value A_1, which may be higher or lower than a_1. The portfolio's original value, denoted $B(a_1)$, now changes to $B(A_1)$. In terms of the horizon H, its new (future) value, denoted $B_H(A_1)$, is:

$$B_H(A_1) = B(A_1) \exp(A_1 H). \tag{49}$$

2.1. The continuously compounded horizon rate of return

The continuously compounded H-horizon rate of return, r_H, transforms an initial investment value, $B(a_1)$, which undergoes an instantaneous change in the rates of interest, into a future value at horizon H. Using (49), r_H is such that

$$B(a_1)\mathrm{e}^{r_H H} = B_H(A_1) = B(A_1)\mathrm{e}^{A_1 H} \tag{50}$$

and hence

$$r_H(A_1) = \frac{1}{H} \log\bigl[B(A_1)/B(a_1)\bigr] + A_1. \tag{51}$$

The horizon rate of return, given a_1, is a function of the two variables A_1 and H. It can thus be viewed as a family of curves in (r_H, A_1) space where H plays the role of a parameter. Those parametric curves are written $r_H = r_H(H; A_1)$.

Before rushing to a diagram representing this family of curves, let us think about the first messages conveyed by formula (51). One of the most immediate is quite remarkable: if interest rates do not change, the horizon rate of return is *independent* from the horizon. Indeed, set $A_1 = a_1$ in (51); r_H then equals a_1 and is independent of H. So we already know that all curves $r_H(H; A_1)$ will go through the fixed point (a_1, a_1). This is not at all obvious: think of 8.75 percent coupon bond, with maturity 22.3 years; consider a horizon of 6 years. It is not evident that if the spot rate remains at 7.25 percent the 6-year horizon rate of return will be 7.25. Only a lengthy reasoning through arbitrage on the bond's value would provide that simple property.

The second message provided by (51) is that when H tends to infinity, the horizon rate of return becomes identical to A_1, i.e., $\lim_{H\to\infty} r_H = A_1$. In (r_H, A_1) space, r_H became the linear function A_1. That has an immediate interpretation: whatever initial capital gain or loss the bond has incurred because a_1 moved to A_1, in the long run only rate A_1 prevails as a determinant of the investor's return.

2.2. A geometrical representation of the horizon rate of return

Now we may want to think of the nature of the curves we must get. We already know that all curves will go through point (a_1, a_1). Additionally, we may surmise that short

horizon curves must be decreasing. The reason is that in the short run capital gains due to a decrease in a_1 far outweigh the losses entailed by coupon reinvestment at a lower rate, and the converse is true in the case of increasing rates. Also, long horizon curves must be increasing, for symmetrical reasons.

We may well venture to guess how any of those curves moves when parameter H increases. Definitely, it must move counter-clockwise. The reason is subtle, and is two-fold. Think how a 2-year horizon curve must behave with regard to a 1-year horizon curve. For values of A_1 smaller than a_1, the 2-year curve must be lower than the 1-year curve: first, the initial capital gain entailed by the spot rate decrease is more eroded after 2 years than after one year; and second, the coupon is reinvested at a lower rate during a longer time. The contrary applies when the rates increase; so definitely the curves will pivot counter-clockwise around the fixed point (a_1, a_1). This is indeed confirmed in Figure 5, where the horizon rate of return is depicted for the holder of a 20-year, 8 percent bond, for horizons equal to 1, 2, 5, 15 years, and for an infinite horizon. The curves look like straight lines, although they are not (with the exception of $r_{H=\infty} = A_1$). Each displays a slight convexity, as we will show later.

$$r_H = \left(\frac{B(a_1)}{B_0}\right)^{1/H} (1 + a_1) - 1$$

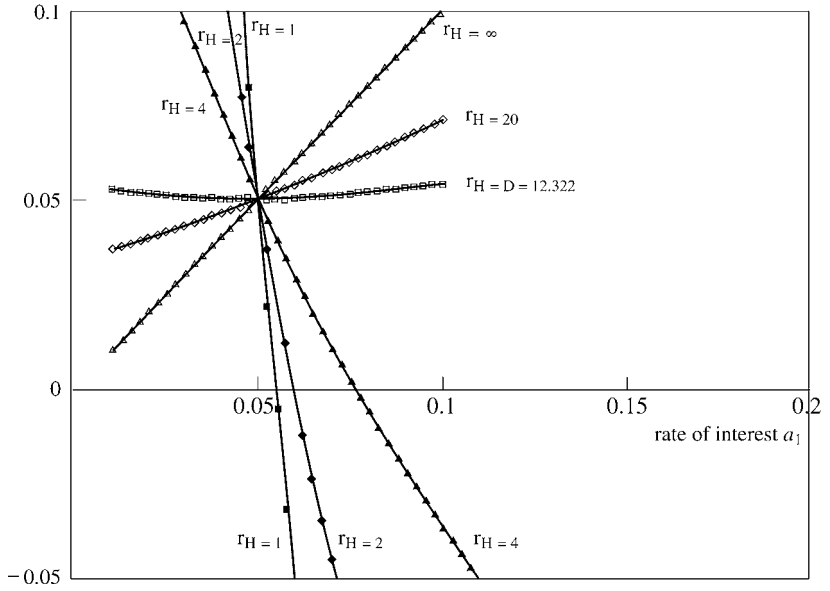


Fig. 5. Horizon rate of return for a bond carrying a 6% coupon rate, with maturity 20 years. Initial rate of interest, when bond is bought: 5%.

The natural question to ask now is whether there exists an intermediate horizon such that the rate of return for that horizon is an (almost flat) curve going through (a_1, a_1). From an economic point of view, it implies that any capital gain (or loss) due to a change in interest rates is compensated by a corresponding loss (or gain) in the reinvestment of coupons. Should this prove to be the case, it would imply that the investor with such a horizon would be protected against any unforeseen change in the interest rates. This horizon exists, and that it corresponds to an essential feature of the bond; we now tackle this question of central importance.

2.3. Existence and characteristics of an immunizing horizon

For such an immunizing horizon to exist, H must be such that r_H goes through a minimum at point (a_1, a_1). A sufficient condition is that $r'_H(a_1) = 0$ and $r''_H(a_1) > 0$. The first-order condition implies

$$r'_H(a_1) = \frac{1}{H}\frac{\mathrm{d}\log B}{\mathrm{d}A_1}\bigg|_{A_1=a_1} + 1 = 0 \tag{52}$$

and hence

$$H = -\frac{\mathrm{d}\log B}{\mathrm{d}A_1}\bigg|_{A_1=a_1}. \tag{53}$$

Thus the horizon should be equal to minus the logarithmic derivative of the bond's value with respect to A_1 at point a_1. Denoting c_t $(t = 1, \ldots, N)$ the cash flows of the bond, we have

$$B(A_1) = \sum_{t=1}^{N} c_t \mathrm{e}^{-A_1 t} \tag{54}$$

and

$$H = -\frac{\mathrm{d}\log B(A_1)}{\mathrm{d}A_1}\bigg|_{A_1=a_1} = -\frac{1}{B(A_1)}\frac{\mathrm{d}B(A_1)}{\mathrm{d}A_1}\bigg|_{A_1=a_1} = \sum_{t=1}^{N} t c_t \mathrm{e}^{-a_1 t}/B(a_1). \tag{55}$$

The last expression in (55) is the weighted average of the times of payment of the bond, the weights being the shares of the present value cash flows in the bond's value. This average is called the *duration* of the bond and is denoted D. It is of central importance in bond immunization. (For a general presentation of this concept and its properties, we take the liberty to refer the reader to our book (de La Grandville, 2001).) Invented by Frederik Macaulay (1938), it bears the name "Macaulay duration" if a is either the spot interest rate, constant for all maturities, or equivalently the bond's yield to maturity. If the cash flows are discounted with a non-constant spot rate structure, it bears the name "Fisher–Weil duration", from Lawrence Fisher and Roman Weil's (1977) paper. Here, the cash flows are discounted with continuous compounding at the constant rate A_1. It implies a discount rate compounded once a year equal to $\exp(A_1) - 1$.

This concept is so central to the immunization process, and possesses such important properties that we will now devote a section to it.

2.4. The Macaulay concept of duration, its properties and uses

After recalling the definition and the properties of the concept of duration, we will outline its uses. The latter are two-fold: first, the concept was invented can be used in immunization of asset against unforeseen changes in a parallel shift of interest rates; second, it is a measure of a bond's or a bond portfolio's sensitivity to such a shift.

Consider first a flat term structure. Call i the rate of interest which, in this case, is unique for all terms. For clarity of exposition and historical exactness, we suppose that i is compounded annually. This implies that the corresponding continuously compounded rate of return is, using our notation, $A_1 = \log(1 + i)$; equivalently, had we wished to express i in terms of A_1, we could have written $i = \exp(A_1) - 1$. The bond value is

$$B(i) = \sum_{t=1}^{T} c_t (1 + i)^{-t}$$

and we are thus led to the following definition.

2.4.1. Definition

The *Macaulay duration* is the weighted average of the bond's times of payment; the weights are the shares of the bond's cash flows in present value in the bond's value. It is thus equal to

$$D = \sum_{t=1}^{T} t c_t (1 + i)^{-t} / B(i).$$

Equivalently, if $B(i)$ is the bond's value observed on the market, i is also the bond's yield to maturity, or the internal rate of return of the investment project of buying the bond at price $B(i)$ and receiving a sum of cash flows in present value equal to $\sum_{t=1}^{T} c_t (1+i)^{-t}$, i.e., yielding a net present value equal to zero. For example, as before, suppose that at time 0 the initial rate is i_0 (continuously compounded, this is, in our notation, $a_1 = \log(1 + i_0)$). At a time ε, very close to zero, suppose that i_0 undergoes a change and becomes a value i (which may be larger or smaller than i_0); using our former notation the new continuously compounded rate of interest is $A_1 = \log(1 + i_0)$.

When the bond was bought, the interest rate was i_0 and we had

$$B(i_0) = \sum_{t=1}^{T} c_t (1 + i_0)^{-t}.$$

This value is exactly equal to that which would have resulted from discounting the cash flows with the initial continuously compounded rate of interest $a_1 = \log(1 + i_0)$,

implying $1 + i_0 = \exp(a_1)$. We have

$$B(i_0) = \sum_{t=1}^{T} c_t (1 + i_0)^{-t} = \sum_{t=1}^{T} c_t \big[\exp(a_1)\big]^{-t} = \sum_{t=1}^{T} c_t \mathrm{e}^{-a_1 t} = B(a_1)$$

which matches (55), where A_1 replaces a_1.

Consider now the horizon rate of return compounded once a year (instead of being continuously compounded as in Section 2.2). It is the rate of return that transforms $B(i_0)$ into its future value after a period H. This future value is $B(i)(1 + i)^H$. Denoting this rate ρ_H, we have

$$B(i_0)(1 + \rho_H)^H = B(i)(1 + i)^H$$

and so ρ_H equals

$$\rho_H(i) = \left[\frac{B(i)}{B(i_0)}\right]^{1/H} (1 + i) - 1.$$

By writing the horizon rate of return $\rho_H(i)$ we are stressing that it is a function of the two variables H and i. If the interest rate does not change and stays fixed at i_0, $\rho_H(i_0) = i_0$.

Consider now the problem of determining a horizon H such that $\rho_H(i)$ stays always above a minimum level of i_0, irrespective of any shock received by the interest translating into a new value i. Minimizing the function $r_H(i)$ is tantamount to minimizing any positive transformation of it; such successive transformations are the following: add 1 to the above expression; take the Hth power, and multiply by $B(i_0)$. We obtain $B(i)(1 + i)^H$, which is the bond's future value at time ε, denoted F_H. Immunizing the horizon rate of return is equivalent to immunizing the bond's future value *for any value* i taken by the rate of interest at time ε. A final transformation is to take the log of the future value. We take the derivative of

$$\log F_H = \log B(i) + H \log(1 + i)$$

with respect to i and equate this to zero at point $i = i_0$ to yield

$$\left.\frac{\mathrm{d} \log F_H}{\mathrm{d}i}\right|_{i=i_0} = \frac{B'(i_0)}{B(i_0)} + \frac{H}{1 + i_0} = 0$$

thus obtaining the candidate solution for the immunization horizon

$$H = -(1 + i_0)\frac{B'(i_0)}{B(i_0)}.$$

This result makes sense from a measurement unit point of view because the left-hand side of the above expression is in years, and so must be the right-hand side. That this is true is seen as follows: $1+i_0$ is unitless because i_0 (expressed in $1/t$) has been multiplied by one year; $B'(i) = \mathrm{d}B/\mathrm{d}i$ is in \$ divided by $1/t$, so these are \$ times t units. Since

$\mathrm{d}B/\mathrm{d}i$ is divided by B (expressed in \$), the result is expressed in $(\$ \cdot t)/\$ =$ time units, as it should.

Now calculate the expression $-(1+i)B'(i)/B(i)$. From

$$B(i) = \sum_{t=1}^{T} c_t (1+i)^t$$

we obtain

$$-(1+i_0)\frac{B'(i_0)}{B(i_0)} = \sum_{t=1}^{T} t c_t (1+i_0)^{-t}/B(i_0) = D.$$

Our result implies that the immunization horizon should equal the Macaulay duration.

The Macaulay duration and the duration which we defined before, using a continuously compounded rate of interest, are one and same concept; they share the *same* value. This stems from the correspondence between i and a_1, which is $i_0 = \exp(a_1) - 1$. Indeed, plug this into the above expression; as shown before, $B(i) = B(A_1)$ and we get

$$\sum_{t=1}^{T} t c_t (1+i)^{-t}/B(i) = \sum_{t=1}^{T} t c_t \big[\exp(a_1)\big]^{-t}/B(a_1) = \sum_{t=1}^{T} t c_t \mathrm{e}^{-a_1 t}/B(a_1)$$

which is our former expression of duration in (55), where cash flows are discounted at a continuously compounded rate.

Another way of confirming that duration has the same value whether we use i or the continuously compounded rate A_1 is to establish directly the equality

$$-\frac{B'(A_1)}{B(A_1)} = -(1+i)\frac{B'(i)}{B(i)}.$$

Consider $B(A_1)$ as the composite function $B[i(A_1)]$. With $i = \exp(A_1) - 1$ and $1 + i = \exp A_1$, we can write

$$B'(A_1) = B'(i) \cdot i'(A_1) = B'(i)\exp(A_1) = B'(i)(1+i),$$

confirming the above equality.

The duration of a bond would be exactly the same if the interest rate used for discounting its cash flows were computed on any other basis, i.e., if they were, for instance, compounded m times per year. Suppose that i_m designates such a yearly rate of interest compounded m times per year; let i stand for i_1, the ordinary rate compounded once a year. We would have

$$1 + i = \left(1 + i_m \cdot \frac{1}{m}\right)^m$$

and

$$\lim_{m\to\infty} \left(1 + i_m \cdot \frac{1}{m}\right)^m = \mathrm{e}^{i_\infty} \equiv \mathrm{e}^{A_1}$$

since A_1 is defined as i_∞.

The values of the bond and of its duration remain unchanged if we discount the cash flows with i_m, whose relationship with i and A_1 can be immediately determined from the above equations as follows:

$$i_m = m\big[(1+i)^{1/m} - 1\big]$$

and

$$i_m = m\big(\mathrm{e}^{A_1/m} - 1\big).$$

The reader may well ask which method of expressing interest and discount rates is preferable. By far the most practical, either in theoretical or applied work, is continuously compounding. Indeed, all of the above and subsequent analysis is valid when time t is not necessarily an integer; and formulas for non integer years of payment are cumbersome when interest and discount rates are compounded once or m times per year; on the other hand, converting those into continuously compounded rates is immediate as shown above.

2.4.2. Properties of the Macaulay duration

Since duration is a weighted average of the times of the bond's payments, it turns out that it can be most usefully seen as a center of gravity. Consider a weightless ruler marked with the times of payment. Load each of those times of payment with the present value cash flows (the measurement unit of these cash flows is arbitrary). Duration will be the center of gravity of these weights (or of the weighted ruler). From this physical representation we can deduce properties 1–3 of duration without recourse to any calculation (a formal proof of property 3 would be quite tedious). All four properties given next are essential inasmuch as duration is a measure of the interest rate risk of a bond.

2.4.2.1. Property 1. Duration and coupon size: Duration decreases with an increase in the coupon. This is obvious since an increase in the coupon will increase all the weights on our ruler in the same proportion, except the last one because it includes the repayment of the principal, which is independent of the coupon. Therefore the center of gravity is displaced to the left; the bond's duration decreases.

2.4.2.2. Property 2. Duration and rate of interest: Duration decreases as the common rate of interest rises. We can observe that an increase in A_1 or in i tends to reduce relatively more the long cash flows than the short ones (because of compounding) and therefore the center of gravity (the duration) will definitely decrease.

2.4.2.3. Property 3. Duration and maturity: Duration does not necessarily increase with the maturity of the bond. This is a surprising result, little known by professionals. Indeed, our intuition may lead us to think that increasing the number of cash flows by one unit will automatically increase duration. This is not so. As pointed out earlier, mathematical analysis of the expression $\mathrm{d}D/\mathrm{d}T$ enables to show that duration does not

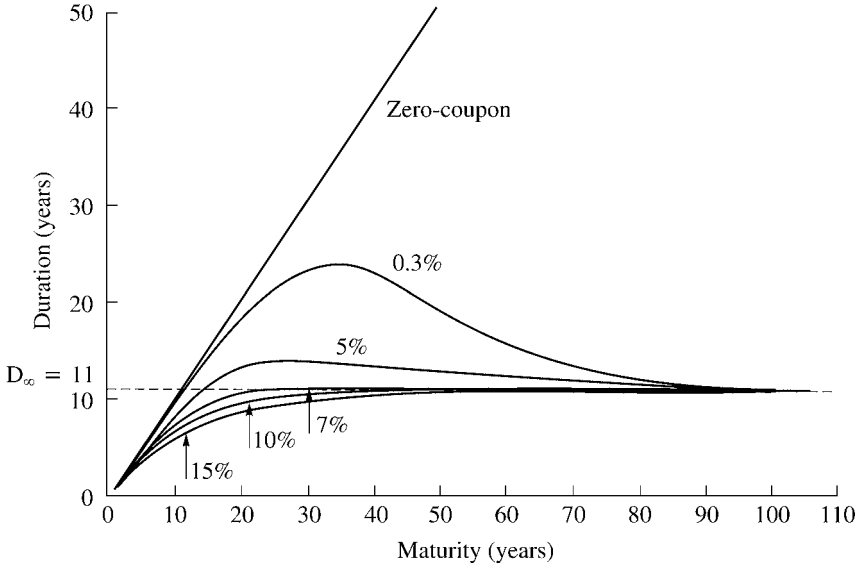


Fig. 6. Duration as a function of maturity. Duration does not necessarily increase with maturity; and for long maturities, duration is almost constant. Source: O. de La Grandville, *Bond Pricing and Portfolio Analysis—Protecting Investors in the Long Run*, MIT Press, 2001, p. 84.

increase necessarily with maturity, but this analysis is quite tedious. However, carefully thinking of duration as a center of gravity enables not only to explain this strange phenomenon, but also to understand under what type of circumstances duration may *decrease* (and therefore the interest rate riskiness of the bond may be *reduced* for longer bonds), all other things remaining equal.

Increasing maturity by one unit has the following consequences. They are summarized in Figure 6. For *high-coupon bonds* (bonds such that their coupon rate is equal or higher than the rate of interest, duration *always* increases with maturity (although at a rate which is both lower than 1, and diminishing with maturity). For *low-coupon bonds*, duration first increase with maturity, similarly as in the high-coupon case; however, it can be shown that there always exists a maturity level beyond which duration *decreases* when maturity increases; typically, this occurs in the vicinity of 25 years.

This apparently surprising result is best understood by considering factors that are liable to displace the center of gravity to the *left*. The fact that maturity increases by one unit has two consequences: first, a new coupon is added on the right-hand side of the rule; second, the weight corresponding to the reimbursement of the principal is *displaced* to the right (by one unit), and at the same time it is *decreased* by a factor $\exp(-A_1)$, or $(1+i)^{-1}$. The right-hand side will be alleviated if the principal is decreased by a sufficiently large factor $\exp(-A_1)$ or $(1+i)^{-t}$ and if the new lever on which the final weight exerts its pressure is relatively far from the initial location of the

center of gravity. Altogether, these factors imply a small coupon, a high rate of interest and a high maturity.

Even if duration does not diminish when maturity increases (which means that the interest rate risk of the bond *decreases* with maturity) Figure 6 shows that there is a wide range of maturities for which duration is almost flat. This has important consequences in the immunization process, as we shall see at the end of Section 2.

Duration tends toward a finite limit when maturity increases infinitely. This means that *all* consols (bonds without reimbursement of the principal, with an infinite stream of coupon payments) have a common duration—a common interest rate risk—whatever the size of the annual coupon. This is also quite surprising. It can be shown as follows. Denoting B_T the bond's reimbursement (par) value, duration can be expressed in closed-form as follows:

$$D = 1 + \frac{1}{i} + \frac{T(i - c/B_T) - (1 + i)}{(c/B_T)[(1 + i)^T - 1] + i}$$

(for a demonstration, see de La Grandville (2001, pp. 80–81)). The third term of the above expression obviously tends to zero when $T \to \infty$. Then

$$\lim_{T \to \infty} D = 1 + \frac{1}{i}.$$

For example, if $i = 5\%$, duration will be $1 + 1/0.05$ years $= 21$ years, irrespective of the coupon size.

This result can be explained as follows. Duration can be thought as the average time the investor has to wait to receive what is owed to him. The inverse of the rate of interest is the time required to receive as interests 100% of any sum loaned. Therefore the investor's average waiting time is $1/i$ *plus* one year since he starts recouping his investment after one year only. This economic interpretation can be generalized: if a coupon is paid m times a year, $\lim_{T \to \infty} D = 1/m + 1/i$, and for a bond paying a continuous coupon (which implies $m \to \infty$), $\lim_{T \to \infty} D = 1/i$.

2.4.2.4. Property 4. Duration of a bond portfolio: *The duration of a bond portfolio is the weighted average of the durations of the bonds comprising the portfolio, the weights being the shares of each bond's value in the portfolio*. The demonstration will be made in Section 3.6.3, in the more general context of a non-flat spot rate structure.

2.5. A second-order condition

Thus a first-order condition for an investment in bonds to be immunized against changes in the interest rates is that the horizon is equal to the duration of the bond. This property will prove essential in the immunization process.

Let us now check whether duration as an investment horizon meets a second-order condition for immunizing our bond. We will revert here to the continuously compounding framework; the analysis with once-a-year compounded interest is the same.

Taking the derivative of (52) with respect to A_1 gives

$$r''_H = \frac{\mathrm{d}}{\mathrm{d}A_1}\left[\frac{1}{H}\frac{B'}{B}\right] = \frac{1}{H}\left[\frac{B''B - B'^2}{B^2}\right] = \frac{1}{H}\left[\frac{B''}{B} - \left(\frac{B'}{B}\right)^2\right]. \tag{56}$$

Evaluating $B''(A_1)/B(A_1)$, and denoting $B(A_1)$ as B, we have

$$B''/B = \frac{1}{B}\frac{\mathrm{d}}{\mathrm{d}A_1}\left[-\sum_{t=1}^{N} t c_t \mathrm{e}^{-A_1 t}\right] = \sum_{t=1}^{N} t^2 c_t \mathrm{e}^{-A_1 t}/B(A_1), \tag{57}$$

which is a measure of the convexity of the bond.

Plugging this result into (56) and using the simplifying notation $c_t \mathrm{e}^{-A_1 t}/B = w_t$, we can write r''_H as

$$r''_H = \frac{1}{H}\left[\sum_{t=1}^{N} t^2 w_t - \left(\sum_{t=1}^{N} t w_t\right)^2\right]. \tag{58}$$

It is useful to check the units of our formula (58). First, r_H is in (1/year), so $r'_H = \mathrm{d}r_H/\mathrm{d}A_1$ is unitless since A_1 is in (1/year). Therefore $r''_H = \mathrm{d}r'_H/\mathrm{d}A_1$ must be in $1/(1/\text{year}) = \text{years}$. This is the case: the right-hand side of (58) is in $(\text{years}^2/\text{year}) = \text{years}$.

The bracketed term is the variance of the times of payment, which is always positive. Thus the second-order condition for a global minimum of $r_H = r_D$ is met. This is confirmed in our example. Our bond has a duration $D = \sum_{t=1}^{N} t c_t \mathrm{e}^{-a_1 t}/B(a_1) = 11.89$ years. The curve $r_{H=D} = 11.89$ goes through a minimum at point a_1. The global convexity of r_H enables to confirm that all curves in Figure 5 are strictly convex (except $r_{H=\infty} = A_1$).

These results immediately extend from a given bond to a bond portfolio, because the formula $\sum_{t=1}^{N} c_t w_t$ for the value of a bond immediately encompasses that of a bond portfolio: c_t stands for any fixed income received at time t, and it may well correspond to a bond portfolio.

2.6. The immunization problem

Until now we have considered the problem of finding a horizon H such that a bond, or a bond portfolio, would be immune to variations of the horizontal spot structure. We can now ask the *converse question*: given a horizon H, is it possible to *build* a portfolio such that its horizon-H rate of return (equivalently: its future value) would be immune to a variation in the rate of interest. The answer is yes: it will always be possible to build such a portfolio; we can even show that there will be an infinite number of solutions. Indeed, suppose we have L bonds, $L > 2$; each has a duration D, $l = 1, \ldots, L$. Let B_l designate bond l's value; let N_l be the number of bonds l in the portfolio, whose value is P. Let w_l be the share of l's position in the portfolio, i.e., $w_l = N_l B_l / P$. Then we

simply have to determine the shares $w_1, \ldots, w_L$ such that

$$\sum_{l=1}^{L} w_l = 1$$

and

$$\sum_{l=1}^{L} w_l D_l = H.$$

This is a system of 2 equations in L ($L > 2$) unknowns; therefore it has an infinite number of solutions. In this highly simplistic scenario, where interest rates have a flat structure and move either up or down in a parallel way, the manager has therefore many options to create a portfolio such that the weighted average of the durations is equal to the investor's horizon. This is not to say that when interest rates obey a more complicated functional relationship, the manager is more restricted in his choices; as we will see later, if the spot rate structure is a polynomial of order r, surprisingly he will need only $2(r + 1) + 1$ bonds to have an infinity of solutions at his disposal.

Furthermore, observe that immunization is a dynamic process. Indeed, as time passes, the investor's horizon is reduced by the same amount. For instance, after 6 months a 10-year investment horizon is reduced to 9 years and a half. Is the portfolio still protected? It is not: even if interest rates have not changed, the duration of *any* coupon-bearing portfolio never diminishes through time at the same pace as maturity (with the rare, temporary, exceptions of portfolios with short and extremely long bonds—see property 3). Typically, after 6 months the duration of the bond portfolio is reduced by *less* than 6 months, and therefore a readjustment of the portfolio is called for. We can also observe that if the horizon is very long (15–20 years), the duration-maturity curve is almost flat: duration hardly changes when the horizon is reduced by 6 months or one year. Suppose now that interest rates meanwhile have moved down: this implies that the duration of the portfolio will have moved *up*, giving further incentive to the manager for readjusting his portfolio. Of course, it may also turn out that interest rates have moved up, taking down the portfolio's duration and therefore bringing it more or less in line with what is needed to allow for the passage of time.

This constitutes what could be described classical immunization, where sensitivities of a bond's portfolio are considered either with respect to an interest rate common to all maturities, or with respect to a common yield to maturity—assuming, of course, that all bonds are default-free. This problem, however, is too simplistic: even if we keep the hypothesis that the investor can choose default-free bonds, he has to cope with the fact that at any given point of time interest rates are not the same whatever the term.

Simplistic a scenario as it may be, it is useful for clarity to see how it can be generalized in successive steps. Consider the general setting in which the spot rate for maturity t, $s(t)$, is expressed as an $m-1$ order polynomial $s(t) = a_1 + a_2 t + a_3 t^2 + \cdots + a_m t^{m-1}$. The total return is

$$S(t) = s(t)t = a_1 t + a_2 t^2 + \cdots + a_m t^m$$

and the values of the portfolio would be a function of the vector $(\mathbf{a} = a_1, \ldots, a_m)$ instead of the scalar a_1. The classical immunization scheme presented just before corresponds to the case where $a_2 = a_3 = \cdots = a_m = 0$; so only the constant a_1 remains: we have $s(t) = a_1$, and $S(t) = s(t)t = a_1 t$. Suppose now that we retain all m coefficients of the polynomial $s(t)$, and that, among the elements of this vector, a_1 only is allowed to move: this implies that the spot rate curve $s(t)$ is shifted by a constant, i.e., it moves parallel to itself.

Since only a_1 is allowed to move to a new value A_1, the horizon rate is still given by (51); the only difference is that in the right-hand side of (51) we now have

$$B(A_1) = \sum_{t=1}^{N} c_t \mathrm{e}^{-(A_1 t + a_2 t^2 + \cdots + a_m t^m)} \tag{59}$$

and

$$B(a_1) = \sum_{t=1}^{N} c_t \mathrm{e}^{-(a_1 t + a_2 t^2 + \cdots + a_m t^m)}. \tag{60}$$

Equation (52) again applies; as to minus the logarithmic derivative of $B(A_1)$ in (53), it is given by

$$-\frac{\mathrm{d}\log B}{\mathrm{d}A_1}\bigg|_{A_1=a_1} = \sum_{t=1}^{N} t c_t \mathrm{e}^{-(a_1 t + a_2 t^2 + \cdots + a_m t^m)} / B(a) \tag{61}$$

where the expression on the right-hand side of (61) is defined as the duration of the bond portfolio.

There is no reason however why such simple scenarios would apply in reality: if considered as a polynomial, there is no reason why the spot rate structure should shift by a constant. Quite on the contrary, short rates are more volatile than longer ones. In consequence a broader analysis is called for, allowing for arbitrary changes in the spot rate structure. This is the question we address in Section 3.

3. Protecting investors against any shift in the interest rate structure—A general immunization theorem

We determine the present and future values of a bond portfolio before and after the shift in the interest rate structure. We then present some new tools which are of central importance in the immunization process: the moment of order k of a bond and of a bond portfolio, as well as fundamental properties of those moments. We will then be ready to state and prove our immunization theorem.

3.1. Notation

We use the following notation:

$L \equiv$ number of different securities (bonds) in the bond portfolio bought at time 0; the bonds are labelled $l = 1, \ldots, L$.

$n_l \equiv$ number of type l bonds in portfolio. The total number of securities in the portfolio is $\sum_1^L n_l$.

$c_{lt} \equiv$ nominal cash flow received by holder of bond l at time t; in monetary units (for instance \$).

$c_t = \sum_{l=1}^L n_l c_{lt} \equiv$ total nominal cash flows received by holder of portfolio at time t; in \$.

$s(t) \equiv$ the continuously compounded interest rate (per year) applying today, at time $t = 0$, for a loan to be reimbursed in t years (t not necessarily an integer). Expressed in (1/year). It is also called the continuously compounded *spot* rate at time 0 for maturity t.

$f(u) \equiv$ the forward rate agreed upon at time 0 for lending at time u ($u > 0$) for an infinitesimally short interval—equivalently: for "instantaneous" lending. Expressed in (1/year).

$S(t) \equiv$ the continuously compounded *total* return over a period t. It is equal to the integral sum of the forward rates over $[0, t]$, $\int_0^t f(u)\,\mathrm{d}u = s(t)t$. It is a pure (dimensionless) number.

$\mathbf{a} = (a_1, a_2, \ldots, a_m) \equiv$ (line) vector of order m of Taylor's series expansion coefficients for $s(t)$ at time 0 when the portfolio is bought; $s(t) \approx a_1 + a_2 t + \cdots + a_m t^{m-1}$. The measurement units of a_j: they are $1/(\text{year})^j$, $j = 1, \ldots, m$. This observation will prove to be quite useful later in this Section. For the time being, observe that $s(t) \approx a_1 + a_2 t + \cdots + a_m t^{m-1}$ is therefore expressed in (1/year) as it should.

$\mathbf{t}^p = (t^1, t^2, \ldots, t^m) =$ (column) vector of m successive positive integer powers of t (the small "p" as a subscript of t stands as "powers").

$S(t) = s(t)t = \int_0^t f(u)\,\mathrm{d}u$ can be approximated by the following polynomial:

$$S(t) \approx a_1 t + a_2 t^2 + \cdots + a_m t^m = \mathbf{a} \cdot \mathbf{t}^p$$

and is expressed as a pure number.

$B_0^l(\mathbf{a}) \equiv$ value of bond l at time 0 (in \$).

$P_0(\mathbf{a}) \equiv$ value of portfolio at time 0 (in \$).

$H \equiv$ immunization horizon, in years; not necessarily an integer number.

$K \equiv$ a positive constant; its dimension is $(\text{years})^{2m}$.

$\mathbf{h}^p = (H^1, H^2, \ldots, H^m) \equiv m$-component vector of successive positive integer powers of horizon H.

$\mathbf{h}_{2m+1}^p \equiv (2m+1)$-component vector, equal to: $(1, H, H^2, \ldots, H^{2m} + K)$.

$\varepsilon \equiv$ time, infinitesimally close to 0, at which the spot rate structure undergoes a variation.

$\mathbf{A} = (A_1, \ldots, A_m) \equiv$ new vector of coefficients of $s(t)$'s Taylor expansion at time ε.

$B^{(l)}(\mathbf{A}) \equiv$ value of bond l at time ε, when the spot rate function $s(t)$ has undergone a variation.

$P_\varepsilon(\mathbf{A}) \equiv$ value of portfolio at time ε, when the $s(t)$ function has undergone a variation.

$B_H^{(l)}(\mathbf{A}) \equiv$ value of bond l at time ε in terms of year H dollars, when the $s(t)$ function has undergone a variation (future value of bond l at time ε).

$P_H(\mathbf{A}) \equiv$ value of portfolio at time ε in terms of year H dollars, when the $s(t)$ function has undergone a variation (future value of portfolio at time ε).
$N \equiv$ portfolio's maturity (in years); maximum maturity of bonds in the portfolio.
$r \equiv$ yield to maturity of bond portfolio; r is defined by the equality $P_0 = \sum_{t=1}^{N} c_t \exp(-rt)$.
$\sum_{t=1}^{N} t c_t \exp(-rt)/P_0 \equiv$ Macaulay duration of portfolio, evaluated with continuous compounding.
$\sum_{t=1}^{N} t c_t \exp[-s(t)t] \equiv$ Fisher–Weil duration $= \sum_{t=1}^{N} t c_t \exp(-\int_0^t f(u)\,\mathrm{d}u) = \sum_{t=1}^{N} t c_t \exp[-S(t)] = \sum_{t=1}^{N} t c_t \exp(-\mathbf{a}\cdot\mathbf{t}^p)$.
We consider, for bonds and bond portfolios, basically four types of value:

- Present value at time $t = 0$, the time of purchase (or in \$ of year 0).
- Present value at time $t = \varepsilon$, an infinitesimally small instant after $t = 0$ (in \$ of year 0), when the spot rate structure has undergone a variation.
- Future value at time $t = 0$ in terms of H-horizon dollars. One should be very clear about the expression "future value", which can be misleading. It does *not* designate the value that the investment will have at time H. It is the amount one would get at $t = H$ by investing at time $t = 0$ the bond price at the spot rate $s(H)$ observed at $t = 0$.
- Future value at time ε in terms of H-horizon dollars: the amount one would get at time H by investing at time ε (or, equivalently, at time 0 with the new structure) the bond price at the *new* spot rate. Both values are equivalent because ε is infinitesimally close to zero.

3.2. Present values at time 0

The present values, at time 0, are:

- for bond l ($l = 1, \ldots, L$):

$$B_0^l(\mathbf{a}) = \sum_{t=1}^{N} c_{lt} \mathrm{e}^{-s(t)t} = \sum_{t=1}^{N} c_{lt} \mathrm{e}^{-\int_0^t f(u)\,\mathrm{d}u} = \sum_{t=1}^{N} c_{lt} \mathrm{e}^{-S(t)}$$
$$= \sum_{t=1}^{N} c_{lt} \mathrm{e}^{-(a_1 t + a_2 t^2 + \cdots + a_j t^j + \cdots + a_m t^m)} = \sum_{t=1}^{N} c_{lt} \exp(-\mathbf{a}\cdot\mathbf{t}^p). \tag{62}$$

- for the portfolio:

$$P_0(\mathbf{a}) = \sum_{t=1}^{N} c_t \mathrm{e}^{-s(t)t} = \sum_{t=1}^{N} c_t \mathrm{e}^{-\int_0^t f(u)\,\mathrm{d}u} = \sum_{t=1}^{N} c_t \mathrm{e}^{-S(t)}$$
$$= \sum_{t=1}^{N} c_t \exp(-\mathbf{a}\cdot\mathbf{t}^p) = \sum_{t=1}^{N} \sum_{l=1}^{L} n_l c_{lt} \exp(-\mathbf{a}\cdot\mathbf{t}^p). \tag{63}$$

3.3. Future values at time 0

The bonds' and portfolios' "future values", expressed at time 0 in terms of H-year dollars, are, respectively:

$$\begin{aligned} B^l_{0,H}(\mathbf{a}) &= B^l_0(\mathbf{a})\mathrm{e}^{s(H)H} = B^l_0(\mathbf{a})\mathrm{e}^{\int_0^H f(u)\,\mathrm{d}u} = B^l_0(\mathbf{a})\mathrm{e}^{S(H)} \\ &= \sum_{t=1}^{N} c_{lt}\mathrm{e}^{-(a_1 t+\cdots+a_j t^j+\cdots+a_m t^m)}\mathrm{e}^{a_1 H+\cdots+a_j H^j+\cdots+a_m H^m} \\ &= \sum_{t=1}^{N} c_{lt}\exp(-\mathbf{a}\cdot\mathbf{t}^p)\cdot\exp(\mathbf{a}\cdot\mathbf{h}^p) \\ &= \sum_{t=1}^{N} c_{lt}\exp\bigl[-\mathbf{a}\cdot(\mathbf{t}^p-\mathbf{h}^p)\bigr]. \end{aligned} \tag{64}$$

Similarly, the portfolio's value in terms of horizon-H years is

$$P_{0,H} = \sum_{t=1}^{N} c_t\exp\bigl[-\mathbf{a}\cdot(\mathbf{t}^p-\mathbf{h}^p)\bigr] = \sum_{t=1}^{N}\sum_{l=1}^{L} n_l c_{lt}\exp\bigl[-\mathbf{a}\cdot(\mathbf{t}^p-\mathbf{h}^p)\bigr]. \tag{65}$$

3.4. Present values at time ε

Suppose that at time ε the spot rate function $s(t)$ has undergone a variation (which may be denoted $\eta(t)$), such that its new approximation as a Taylor expansion is

$$s_\varepsilon(t) = A_1 + A_2 t + \cdots + A_j t^{j-1} + \cdots + A_m t^{m-1}. \tag{66}$$

Then

$$S_\varepsilon(t) = A_1 t + A_2 t^2 + \cdots + A_j t^j + \cdots + A_m t^m = \mathbf{A}\cdot\mathbf{t}^p. \tag{67}$$

The new present and future values formulas for the bond and the bond portfolio keep the same form as above, except that vector $\mathbf{A}$ now replaces vector $\mathbf{a}$; the present values are

$$B^l_\varepsilon(A) = \sum_{t=1}^{N} c_{lt}\exp(-\mathbf{A}\cdot\mathbf{t}^p) \tag{68}$$

and

$$P_\varepsilon(A) = \sum_{t=1}^{N} c_t\exp(-\mathbf{A}\cdot\mathbf{t}^p) = \sum_{t=1}^{N}\sum_{l=1}^{S} n_l c_{lt}\exp(-\mathbf{A}\cdot\mathbf{t}^p). \tag{69}$$

3.5. Future values at time ε

Under the new spot rate structure, the future values are, at time ε

$$B^l_{\varepsilon,H}(A) = \sum_{t=1}^{N} c_{lt} \exp\bigl[-\mathbf{A}\cdot(\mathbf{t}^p - \mathbf{h}^p)\bigr] \tag{70}$$

and

$$P_{\varepsilon,H}(A) = \sum_{t=1}^{N} c_t \exp\bigl[-\mathbf{A}\cdot(\mathbf{t}^p - \mathbf{h}^p)\bigr] = \sum_{t=1}^{N}\sum_{l=1}^{S} n_l c_{lt} \exp\bigl[-\mathbf{A}\cdot(\mathbf{t}^p - \mathbf{h}^p)\bigr]. \tag{71}$$

3.6. Further concepts for immunization: the moments of order k of a bond and a bond portfolio

We now introduce some little known, but essential, concepts: the moment of order k of a bond or a bond portfolio. This concept is very much akin to the moment of a random variable, with two main differences: the objects for which the calculation of moments are made are not random, but deterministic, and their values are always positive.

3.6.1. The moment of order k of a bond

Definition. The *moment of order k of a bond* is the weighted average of the kth power of its times of payments, the weights being the shares of the bond's cash flows in present value in the bond's present value. For bond l, this moment of order k is

$$\mu^{(l)}_k = \sum_{t=1}^{N} t^k c_{lt} \mathrm{e}^{-s(t)t}/B^l_0 = \sum_{t=1}^{N} t^k c_{l_t} \mathrm{e}^{-\int_0^t f(u)\,\mathrm{d}u}/B^l_0 = \sum_{t=1}^{N} t^k c_{l_t} \mathrm{e}^{-S(t)}/B^l_0,$$
$$k = 0, 1, 2, \ldots,$$

and, using our assumption that about $s(t)$:

$$\mu^{(l)}_k = \sum_{t=1}^{N} t^k c_{lt} \exp(-\mathbf{a}\cdot\mathbf{t}^p)/B^l_0(a), \quad k = 0, 1, 2, \ldots. \tag{72}$$

A moment of order k is expressed in (year)^k units.

3.6.2. The moment of order k of a bond portfolio

Definition. The moment of order k of a bond portfolio is the weighted average of the kth power of its times of payments, the weights being the shares of the portfolio's cash

flows in present value in the portfolio's present value. Denoted $\mu_k^{(P)}$, it is equal to:

$$\mu_k^{(P)} = \sum_{t=1}^{N} t^k c_t \exp(-\mathbf{a} \cdot \mathbf{t}^p)/P_0(a), \quad k = 0, 1, 2, \ldots. \tag{73}$$

The moment of order 0 of a portfolio (or a bond) is equal to 1, since it is the weighted average of 1's. Also, the moment of order one of a portfolio (or a bond) is its Fisher–Weil duration.

3.6.3. Properties of moments

We first show the relationship between the moments of order k of a portfolio and the moments of same order of the bonds making up the portfolio. We then show how these moments measure sensitivities of the bond prices to changes in the coefficients of the $s(t)$'s polynomial.

Theorem 3. *The kth moment of a bond portfolio is the weighted average of the kth moments of the bonds making up the portfolio*; *the weights are the shares of each bond in the portfolio.*

Proof. From (73) we can write successively:

$$\begin{aligned}
\mu_k^{(P)} &= \sum_{t=1}^{N} t^k c_t \exp(-\mathbf{a} \cdot \mathbf{t}^p)/P_0(\mathbf{a}) = \sum_{t=1}^{N} t^k \sum_{l=1}^{L} n_l c_{lt} \exp(-\mathbf{a} \cdot \mathbf{t}^p)/P_0(\mathbf{a}) \\
&= \sum_{l=1}^{L} \sum_{t=1}^{N} t^k n_l c_{lt} \exp(-\mathbf{a} \cdot \mathbf{t}^p)/P_0(\mathbf{a}) \\
&= \sum_{l=1}^{L} n_l \sum_{t=1}^{N} t^k c_{lt} \exp(-\mathbf{a} \cdot \mathbf{t}^p)/P_0(\mathbf{a}) \\
&= \sum_{l=1}^{L} \frac{n_l B_0^l(\mathbf{a})}{P_0(\mathbf{a})} \sum_{t=1}^{N} t^k c_{lt} \exp(-\mathbf{a} \cdot \mathbf{t}^p)/B_0^l(\mathbf{a}).
\end{aligned} \tag{74}$$

If $\alpha_l = \frac{n_l B_0^l(\mathbf{a})}{P_0(\mathbf{a})}$ designates the share of bond l's value in the initial portfolio ($\sum_{l=1}^{L} \alpha_l = 1$), then:

$$\mu_k^{(P)} = \sum_{l=1}^{L} \alpha_l \mu_k^{(l)}, \quad k = 0, 1, 2, \ldots. \tag{75}$$

Theorem 4. *Minus the logarithmic derivative of the l bond's* [*the portfolio's*] *price with respect to coefficient a_j is equal to the jth moment of the bond* [*the portfolio*].

Proof. We prove this theorem for a bond portfolio, a single bond being just a particular case.

From (63):

$$P_0(\mathbf{a}) = \sum_{t=1}^{N} c_t \exp(-\mathbf{a} \cdot \mathbf{t}^p).$$

Its logarithmic derivative (its relative rate of increase in linear approximation) with respect to a_j is

$$\frac{1}{P_0(\mathbf{a})} \frac{\partial P_0(\mathbf{a})}{\partial a_j} = -\sum_{t=1}^{N} t^j c_t \exp(-\mathbf{a} \cdot \mathbf{t}^p)/P_0(\mathbf{a}) \quad (j = 1, \ldots, m). \tag{76}$$

Applying (73), we recognize in the right-hand side of (76) minus the jth moment of the portfolio. Thus we have the fundamental relationship contained in Theorem 4

$$\frac{-1}{P_0(\mathbf{a})} \frac{\partial P_0(\mathbf{a})}{\partial a_j} = \mu_j^{(P)} \quad (j = 1, \ldots, m). \tag{77}$$

Three remarks are in order at this point. First, the dimensions of both sides of (77) do match, as they should: from our observation in Section 3.1 on the measurement units of a_j $(j = 1, \ldots, m)$, the left-hand side of (77) is expressed in $1/[1/(\text{year})^j] = \text{year}^j$. These are indeed the units of the jth moment of the portfolio.

Second, suppose that the polynomial $s(t)$ reduces to a_1. We then have $s(t) = a_1 \equiv r$ (a constant), and $S(t) = s(t)t = \int_0^t f(u)\,\mathrm{d}u = \int_0^t a_1\,\mathrm{d}u = a_1 t$. The portfolio's value reduces to $P_0(a_1) = \sum_{t=1}^{N} c_t \exp(-a_1 t)$; a_1 is nothing else than the portfolio's yield to maturity, equal to r, by the definition of the yield to maturity given in Section 3.1.

We then have

$$\frac{-1}{P_0(t)} \frac{\mathrm{d}P_0(a_1)}{\mathrm{d}a_1} = \sum_{t=1}^{N} t c_t \exp(-a_1 t)/P_0(a_1). \tag{78}$$

This reduces to the well-known property that the logarithmic derivative of the portfolio's value with respect to the constant $a_1 = r$ (or the yield to maturity) is equal to minus its Macaulay duration evaluated with continuous compounding.

Finally, suppose that the spot rate structure does not reduce to a constant, but that it retains its general form of an $(m-1)$th order polynomial. However, consider that only coefficient a_1 moves: this implies that the whole structure (which is not a constant) is shifted up or down by a given amount. In other words, the structure moves in a parallel fashion. Taking the logarithmic partial derivative of (63) with respect to a_1 (or, equivalently, replacing a_j by a_1 in (76))

$$\frac{-1}{P_0(\mathbf{a})} \frac{\partial P_0(\mathbf{a})}{\partial a_1} = \sum_{t=1}^{N} t c_t \exp(-\mathbf{a} \cdot \mathbf{t}^p)/P_0(\mathbf{a}). \tag{79}$$

This time our particular case amounts to the equality between the portfolio's sensitivity to a *parallel* change in the spot rate structure and the Fisher–Weil duration.

3.7. A general immunization theorem

We will now state and prove the following immunization theorem.

Theorem 5. *Suppose that the spot rate structure can be expanded into a Taylor series of order* $m-1$*, and that it undergoes a variation. Then a sufficient condition for a bond portfolio to be immunized against such a variation is the following*:

- *any moment of order* k $(k = 0, 1, \ldots, 2m-1)$ *of the bond portfolio is equal to the* k*th power of the investor's horizon* H *and*
- *the moment of order* $2m$ *is equal to the* $2m$*th power of* H *plus a positive, arbitrary constant.*

Proof and discussion

Suppose that the initial spot rate structure $s(t)$ can be approximated to the $(m-1)$th order by the following polynomial

$$s(t) \approx a_1 + a_2 t + a_3 t^2 + \cdots + a_j t^{j-1} + \cdots + a_m t^{m-1} \tag{80}$$

where $a_j = \frac{1}{(j-1)!} s^{(j-1)}(0)$, $j = 1, \ldots, m$; $s^{(j-1)}(0)$ is the $(j-1)$th derivative of $s(t)$ at $t = 0$.

The total return $S(t) = s(t)t$ is therefore approximated by

$$S(t) = a_1 t + a_2 t^2 + a_3 t^3 + \cdots + a_j t^j + \cdots + a_m t^m = \mathbf{a} \cdot \mathbf{t}^p. \tag{81}$$

The implied forward rate function is

$$f(t) = S'(t) = a_1 + 2a_2 t + 3a_3 t^2 + \cdots + j a_j t^{j-1} + \cdots + m a_m t^{m-1}. \tag{82}$$

(The reader may verify on this example one of the properties we had demonstrated under Section 1.5, namely that at the origin the slope of the forward rate curve is twice that of the spot rate curve. Indeed, at the origin the slope of the spot rate curve is $s'(0) = a_2$, while $f'(0) = 2a_2$.)

The value of a bond portfolio bought at time 0, under the initial spot rate structure, is

$$P_0(a) \approx \sum_{t=1}^{N} c_t \exp(-\mathbf{a} \cdot \mathbf{t}^p).$$

Under the initial spot structure $\mathbf{a} \cdot \mathbf{t}^p$, the portfolio's *future* value, at horizon H is

$$P_{0,H}(a) \approx P_0(a) \mathrm{e}^{a_1 H + a_2 H^2 + \cdots + a_m H^m} = P_0(a) \exp(\mathbf{a} \cdot \mathbf{h}^p)$$

where $\mathbf{h}^p$ denotes the vector $(H, H^2, \ldots, H^m)$. Therefore

$$P_{0,H}(a) = \sum_{t=1}^{N} c_t \exp(-\mathbf{a}\cdot\mathbf{t}^p)\exp(\mathbf{a}\cdot\mathbf{h}^p) = \exp(\mathbf{a}\cdot\mathbf{h}^p)\sum_{t=1}^{N} c_t \exp(-\mathbf{a}\cdot\mathbf{t}^p).$$

Suppose now that the whole spot rate structure $s(t)$ undergoes a variation: from $\mathbf{a}\cdot\mathbf{t}^p$ at time 0 it becomes $\mathbf{A}\cdot\mathbf{t}^p$ at time ε. Therefore the future value of the bond portfolio becomes

$$P_{\varepsilon,H} = \left[\sum_{t=1}^{N} c_t \exp(-\mathbf{A}\cdot\mathbf{t}^p)\right]\exp(\mathbf{A}\cdot\mathbf{h}^p). \tag{83}$$

First order conditions. For the portfolio to be immunized, a first order condition is that the gradient of any positive transformation of $P_{\varepsilon,H}$ equals zero. Taking logs on both sides of (83) yields

$$\log P_{\varepsilon,H} = \mathbf{A}\cdot\mathbf{h}^p + \log\left[\sum_{t=1}^{N} c_t \exp(-\mathbf{A}\cdot\mathbf{t}^p)\right]. \tag{84}$$

Differentiate (84) with respect to the m variables $A_1, \ldots, A_j, \ldots, A_m$, and equate to zero these derivatives at point $\mathbf{a}$:

$$\frac{\partial \log P_{\varepsilon,H}(\mathbf{a})}{\partial A_j} = H^j + \frac{-1}{P_0}\left[\sum_{t=1}^{N} t^j c_t \exp(-\mathbf{a}\cdot\mathbf{t}^p)\right] = 0, \quad j = 1, \ldots, m. \tag{85}$$

These m equations (85) imply that each jth moment of the portfolio must be equal to the jth power of the investor's horizon:

$$\mu_j^{(P)} = H^j, \quad j = 1, \ldots, m. \tag{86}$$

A first set of conditions is defined by the $m + 1$ equations corresponding to:

(a) the m equations implied by (86);
(b) an additional equation corresponding to the accounting constraint: the sum of all shares of each bond in the portfolio must be equal to one.

We will take up these conditions in order, together with their implications in terms of the amounts of the bonds to be chosen.

For (a) to be met, let us use the property that the portfolio's moment of order j is the weighted average of each bond's moment of order j, the weights being the share of each bond in the portfolio

$$\mu_j^{(P)} = \sum_{l=1}^{L} \frac{n_l B_0^l(\mathbf{a})}{P_0(\mathbf{a})}\mu_j^{(l)}, \quad j = 1, \ldots, m. \tag{87}$$

Thus (86) translates as

$$\sum_{l=1}^{L} \frac{n_l B_0^l(\mathbf{a})}{P_0(\mathbf{a})} \mu_j^{(l)} = H^j, \quad j = 1, \ldots, m. \tag{88}$$

Furthermore, the L unknowns n_l ($l = 1, \ldots, L$) must be such that the sum of the shares $n_l B_0^l(\mathbf{a})/P_0(\mathbf{a})$ equals one. In addition to the m equations (88), the L unknowns n_l must verify

$$\sum_{l=1}^{L} n_l B_0^l(\mathbf{a})/P_0(\mathbf{a}) = 1. \tag{89}$$

So Equations (88) and (89) amount to a system of $m + 1$ equations in $m + 1 = L$ unknowns n_l ($l = 1, \ldots, m + 1$). To gain insight into the structure of system (88), (89), it may be useful to write it out completely. Before proceeding, notice that Equation (89) has exactly the same structure as (88): it implies that *all* moments of order j of the portfolio must be equal to the jth power of the horizon H^j, *including* the moment of order zero: that moment is just equal to the sum of the shares of each bond in the portfolio, which must be equal to one.

For clarity, let us express (88) and (89) in terms of their components. This system of $m + 1$ equations can be written in the following way, starting with Equation (89), followed by system (88).

Since the moment of order 0 of each bond is equal to one by definition, Equation (89) and system (88) can be written equivalently as:

$$\begin{aligned}
&\frac{B_0^1(\mathbf{a})}{P_0(\mathbf{a})}\mu_0^{(1)}n_1 + \frac{B_0^2(\mathbf{a})}{P_0(\mathbf{a})}\mu_0^{(2)}n_2 + \cdot + \frac{B_0^l(\mathbf{a})}{P_0(\mathbf{a})}\mu_0^{(l)}n_l + \cdot \\
&\quad + \frac{B_0^{(m+1)}(\mathbf{a})}{P_0(\mathbf{a})}\mu_0^{(m+1)}n_{m+1} = 1, \\
&\frac{B_0^1(\mathbf{a})}{P_0(\mathbf{a})}\mu_1^{(1)}n_1 + \frac{B_0^2(\mathbf{a})}{P_0(\mathbf{a})}\mu_1^{(2)}n_2 + \cdot + \frac{B_0^l(\mathbf{a})}{P_0(\mathbf{a})}\mu_1^{(l)}n_l + \cdot \\
&\quad + \frac{B_0^{(m+1)}(\mathbf{a})}{P_0(\mathbf{a})}\mu_1^{(m+1)}n_{m+1} = H, \\
&\cdots \\
&\frac{B_0^1(\mathbf{a})}{P_0(\mathbf{a})}\mu_k^{(1)}n_1 + \frac{B_0^2(\mathbf{a})}{P_0(\mathbf{a})}\mu_k^{(2)}n_2 + \cdot + \frac{B_0^l(\mathbf{a})}{P_0(\mathbf{a})}\mu_k^{(l)}n_l + \cdot \\
&\quad + \frac{B_0^{(m+1)}(\mathbf{a})}{P_0(\mathbf{a})}\mu_k^{(m+1)}n_{m+1} = H^k, \\
&\cdots
\end{aligned}$$

$$\frac{B_0^1(\mathbf{a})}{P_0(\mathbf{a})}\mu_m^{(1)}n_1 + \frac{B_0^2(\mathbf{a})}{P_0(\mathbf{a})}\mu_m^{(2)}n_2 + \cdot + \frac{B_0^l(\mathbf{a})}{P_0(\mathbf{a})}\mu_m^{(l)}n_l + \cdot$$

$$+ \frac{B_0^{(m+1)}(\mathbf{a})}{P_0(\mathbf{a})}\mu_m^{(m+1)}n_{m+1} = H^m. \tag{90}$$

A word of caution is needed here. It would be tempting to solve the above system of $m+1$ equations in the $m+1$ unknowns $n_1, \ldots, n_{m+1}$ and consider the solution as a possible solution for the immunizing portfolio. This system reflects the first-order conditions only and not the second-order ones. We will show that we can never form an optimal portfolio by relying on first-order conditions only.

Second order conditions. Let us now address second-order conditions and their implications. For a local minimum of $P_H(\mathbf{A})$ at $\mathbf{A} = \mathbf{a}$, to Equations (90) we have to add those corresponding to a positive semi-definite Hessian matrix for $P_H(\mathbf{A})$ at $\mathbf{A} = \mathbf{a}$. We will now tackle these conditions.

Recall that the portfolio's future value at horizon H and at time ε is given by (83), which we can write equivalently as

$$P_{\varepsilon;H}(\mathbf{A}) = \sum_{t=1}^{N} c_t \exp\big[-\mathbf{A}\cdot(\mathbf{t}^p - \mathbf{h}^p)\big]. \tag{91}$$

Taking the partial derivative of (91) with respect to A_j $(j = 1, \ldots, m)$ yields

$$\frac{\partial P_{\varepsilon;H}}{\partial A_j}(\mathbf{A}) = -\sum_{t=1}^{N} c_t(t^j - H^j)\exp\big[-\mathbf{A}\cdot(\mathbf{t}^p - \mathbf{h}^p)\big], \quad j = 1, \ldots, m. \tag{92}$$

The partial derivative of (92) with respect to A_i yields the generic element of the Hessian matrix:

$$\frac{\partial^2 P_{\varepsilon;H}}{\partial A_i \partial A_j}(\mathbf{A}) = \sum_{t=1}^{N} c_t(t^j - H^j)(t^i - H^i)\exp\big[-\mathbf{A}\cdot(\mathbf{t}^p - \mathbf{h}^p)\big],$$
$$i = 1, \ldots, m;\ j = 1, \ldots, m. \tag{93}$$

To simplify notation, let

$$g_t \equiv c_t \exp\big[-\mathbf{A}\cdot(\mathbf{t}^p - \mathbf{h}^p)\big], \quad t = 1, \ldots, N, \tag{94}$$

denote the portfolio's cash flow received at time t expressed in future value (i.e., in horizon H's value). Each g_t depends upon $\mathbf{A}$.

Let Q_{ij} denote the generic element of the Hessian matrix $\partial^2 P_{\varepsilon,H}(\mathbf{a})/\partial A_i \partial A_j$, $i = 1, \ldots, m;\ j = 1, \ldots, m$. I owe to Anthony Pakes the idea of developing Q_{ij} as follows

$$Q_{i,j} = \sum_{t=1}^{N} g_t (t^j - H^j)(t^i - H^i) = \sum_{t=1}^{N} g_t (t^i t^j - t^i H^j - H^i t^j + H^i H^j)$$
$$= \sum_{t=1}^{N} g_t t^{i+j} - H^j \sum_{t=1}^{N} g_t t^i - H^i \sum_{t=1}^{N} g_t t^j + H^{i+j} \sum_{t=1}^{N} g_t,$$
$$i = 1, \ldots, m; \;\; j = 1, \ldots, m. \tag{95}$$

From the first-order conditions (85)

$$\frac{1}{P_0} \sum_{t=1}^{N} t^j c_t \exp(-\mathbf{a} \cdot \mathbf{t}^p) = H^j, \quad j = 1, \ldots, m, \tag{85}$$

we can deduce

$$\sum_{t=1}^{N} t^j c_t \exp(-\mathbf{a} \cdot \mathbf{t}^p) = H^j \sum_{t=1}^{N} c_t \exp(-\mathbf{a} \cdot \mathbf{t}^p) \tag{96}$$

and

$$\sum_{t=1}^{N} t^j c_t \exp\big[-\mathbf{a} \cdot (\mathbf{t}^p - \mathbf{h}^p)\big] = H^j \sum_{t=1}^{N} c_t \exp\big[-\mathbf{a} \cdot (\mathbf{t}^p - \mathbf{h}^p)\big].$$

Therefore

$$\sum_{t=1}^{N} g_t t^j = H^j \sum_{t=1}^{N} g_t, \quad j = 1, \ldots, m,$$

which is also valid when j is replaced by i ($i = 1, \ldots, m$). In expression (95), $\sum_{t=1}^{N} g_t t^i$ and $\sum_{t=1}^{N} g_t t^j$ can thus be replaced by $H^i \sum_{t=1}^{N} g_t$ and $H^j \sum_{t=1}^{N} g_t$ respectively, and the generic term of the Hessian at $\mathbf{a}$ becomes

$$Q_{i,j}(\mathbf{a}) = \sum_{t=1}^{N} g_t t^{i+j} - H^{i+j} \sum_{t=1}^{N} g_t \tag{97}$$

or, reverting to the notation using the c_t's

$$Q_{i,j}(\mathbf{a}) = \sum_{t=1}^{N} c_t \exp\big[-\mathbf{a} \cdot (\mathbf{t}^p - \mathbf{h}^p)\big] t^{i+j}$$
$$- H^{i+j} \sum_{t=1}^{N} c_t \exp\big[-\mathbf{a} \cdot (\mathbf{t}^p - \mathbf{h}^p)\big]. \tag{98}$$

Multiplying and dividing the first term in the right-hand side of Equation (98) by $P_0 = \sum_{t=1}^{N} c_t \exp(-\mathbf{a} \cdot \mathbf{t}^p)$ yields

$$
\begin{aligned}
Q_{i,j}(\mathbf{a}) &= P_0 \sum_{t=1}^{N} t^{i+j} c_t \exp\bigl[-\mathbf{a}\cdot(\mathbf{t}^p - \mathbf{h}^p)\bigr]/P_0 - H^{i+j} P_0 \exp(\mathbf{a}\cdot\mathbf{h}^p) \\
&= P_0 \exp(\mathbf{a}\cdot\mathbf{h}^p)\mu_{i+j}^{(P)} - H^{i+j} P_0 \exp(\mathbf{a}\cdot\mathbf{h}^p) \\
&= P_0 \exp(\mathbf{a}\cdot\mathbf{h}^p)\bigl[\mu_{i+j}^{(P)} - H^{i+j}\bigr] \\
&= P_{0,H}\bigl[\mu_{i+j}^{(P)} - H^{i+j}\bigr], \quad 1 \leqslant i \leqslant m,\ 1 \leqslant j \leqslant m. \qquad (99)
\end{aligned}
$$

Call the difference between the moment of order k and the kth power H the *generalized variance* of order k of the portfolio, and denote this generalized variance by

$$\gamma_k = \mu_k^{(P)} - H^k, \quad k > 0. \qquad (100)$$

All elements $Q_{i,j}(\mathbf{a})$ of the Hessian matrix are those of the matrix $\mu_{i+j}^{(P)} - H^{i+j} = \gamma_{i+j} \equiv \gamma_{i,j}$, each one being multiplied by the positive constant $P_{0,H}$. So the discussion of the positive semi-definiteness of matrix $\mathbf{Q}$ whose elements are $Q_{i,j}$ reduces to that of matrix $\mathbf{\Gamma}_m \equiv (\gamma_{i+j})$.

The first-order conditions imply that all generalized variances up to the mth order must vanish. So all elements in the triangle above the second principal diagonal of the $\mathbf{\Gamma}_m$ matrix are equal to zero. (We call "second principal diagonal" of matrix $\mathbf{\Gamma}_m$ the diagonal between its lower-left element and its upper-right element; this diagonal has m elements, all equal to γ_{m+1}.) All elements in any diagonal parallel to the second diagonal must be equal to one another: indeed $\gamma_{i,j} = \gamma_{i+j}$.

We are now left with the task of looking for the properties that matrix $\mathbf{\Gamma}_m$ must exhibit in order to be positive semi-definite. $\mathbf{\Gamma}_m$ may be written as follows:

$$
\mathbf{\Gamma}_m = \begin{bmatrix}
0 & 0 & 0 & \cdots & \gamma_{m+1} \\
0 & 0 & & \gamma_{m+1} & \gamma_{m+2} \\
 & & & \gamma_{m+2} & \gamma_{m+3} \\
0 & & \vdots & \vdots & \\
\vdots & \gamma_{m+1} & & & \\
\gamma_{m+1} & \gamma_{m+2} & \gamma_{m+3} & \cdots & \gamma_{2m}
\end{bmatrix}.
$$

For $\mathbf{\Gamma}_m$ to be positive semi-definite, all its principal minors must be either positive or zero. Consider, in the above matrix, the (2×2) principal minor exhibiting one zero in its upper-left corner. There are two possibilities:

(i) m is even; the above-mentioned principal minor is then of the form

$$\begin{vmatrix} 0 & \gamma_{m+1} \\ \gamma_{m+1} & \gamma_{m+2} \end{vmatrix}.$$

For this minor to be positive or zero, γ_{m+1} must be zero.

(ii) m is odd; the principal minor is

$$\begin{vmatrix} 0 & 0 \\ 0 & \gamma_{m+1} \end{vmatrix}$$

and, adding one line and one column, an adjacent principal minor is then

$$\begin{vmatrix} 0 & 0 & \gamma_{m+1} \\ 0 & \gamma_{m+1} & \gamma_{m+2} \\ \gamma_{m+1} & \gamma_{m+2} & \gamma_{m+3} \end{vmatrix}$$

which must also be positive or zero. For $\mathbf{\Gamma}_m$ to be positive semi-definite, γ_{m+1} has to be positive or equal to zero. Suppose it is positive. Then this adjacent minor is equal to $-\gamma_{m+1}^3$, which is negative, entailing a contradiction. Thus γ_{m+1} must equal zero.

Let us now take up γ_{m+2}. Consider again case (i). γ_{m+2} being a principal minor, it must be either positive or equal to zero. Suppose it is positive. An adjacent principal minor is

$$\begin{vmatrix} 0 & \gamma_{m+1} & \gamma_{m+2} \\ \gamma_{m+1} & \gamma_{m+2} & \gamma_{m+3} \\ \gamma_{m+2} & \gamma_{m+3} & \gamma_{m+4} \end{vmatrix} = \begin{vmatrix} 0 & 0 & \gamma_{m+2} \\ 0 & \gamma_{m+2} & \gamma_{m+3} \\ \gamma_{m+2} & \gamma_{m+3} & \gamma_{m+4} \end{vmatrix}$$

(from what has been shown above), and its value is $-\gamma_{m+2}^3 < 0$, which is not allowed. So γ_{m+2} must be zero.

In case (ii), a principal minor is

$$\begin{vmatrix} \gamma_{m+1} & \gamma_{m+2} \\ \gamma_{m+2} & \gamma_{m+3} \end{vmatrix} = \begin{vmatrix} 0 & \gamma_{m+2} \\ \gamma_{m+2} & \gamma_{m+3} \end{vmatrix},$$

and its value is $-\gamma_{m+2}^2$. For it to be positive or zero, γ_{m+2} has to be zero.

Carrying on the same argument, all terms γ_k must be equal to zero, except the last one, equal to $\gamma_{2m} = \mu_{2m} - H^{2m}$, which can be set at any (arbitrary) positive value K, expressed in (years)^{2m}.

So the first-order conditions imply the m equations

$$\begin{aligned} &\mu_1^{(P)} = H, \\ &\mu_2^{(P)} = H^2, \\ &\vdots \\ &\mu_m^{(P)} = H^m. \end{aligned}$$

The second-order conditions imply the m following equations

$$\begin{aligned} &\mu_{m+1}^{(P)} = H^{m+1}, \\ &\mu_{m+2}^{(P)} = H^{m+2}, \\ &\vdots \\ &\mu_{2m-1}^{(P)} = H^{2m-1}, \\ &\mu_{2m}^{(P)} = H^{2m} + K \end{aligned}$$

to which finally the accounting constraint $\mu_0 = 1$ must be added, for a total of $2m + 1$ equations, implying $2m + 1$ bonds in the immunizing portfolio. This completes the proof of Theorem 5. As an example, if a polynomial of order 3 is used for the spot rate structure, 4 coefficients correspond to that polynomial ($m = 4$). The immunizing portfolio will then be made out of $2m + 1 = 9$ bonds.

The quantities n_l ($l = 1, \ldots, 2m + 1$) of each bond are the solutions of the following system of $2m + 1$ equations:

$$
\begin{aligned}
&\frac{B_0^1(\mathbf{a})}{P_0(\mathbf{a})}\mu_0^{(1)}n_1 + \frac{B_0^2(\mathbf{a})}{P_0(\mathbf{a})}\mu_0^{(2)}n_2 + \cdot + \frac{B_0^l(\mathbf{a})}{P_0(\mathbf{a})}\mu_0^{(l)}n_l + \cdot \\
&\quad + \frac{B_0^{(2m+1)}(\mathbf{a})}{P_0(\mathbf{a})}\mu_0^{(2m+1)}n_{2m+1} = 1, \\
&\frac{B_0^1(\mathbf{a})}{P_0(\mathbf{a})}\mu_1^{(1)}n_1 + \frac{B_0^2(\mathbf{a})}{P_0(\mathbf{a})}\mu_1^{(2)}n_2 + \cdot + \frac{B_0^l(\mathbf{a})}{P_0(\mathbf{a})}\mu_1^{(l)}n_l + \cdot \\
&\quad + \frac{B_0^{(2m+1)}(\mathbf{a})}{P_0(\mathbf{a})}\mu_1^{(2m+1)}n_{2m+1} = H, \\
&\ldots \\
&\frac{B_0^1(\mathbf{a})}{P_0(\mathbf{a})}\mu_k^{(1)}n_1 + \frac{B_0^2(\mathbf{a})}{P_0(\mathbf{a})}\mu_k^{(2)}n_2 + \cdot + \frac{B_0^l(\mathbf{a})}{P_0(\mathbf{a})}\mu_k^{(l)}n_l + \cdot \\
&\quad + \frac{B_0^{(2m+1)}(\mathbf{a})}{P_0(\mathbf{a})}\mu_k^{(2m+1)}n_{2m+1} = H^k, \\
&\ldots \\
&\frac{B_0^1(\mathbf{a})}{P_0(\mathbf{a})}\mu_{2m}^{(1)}n_1 + \frac{B_0^2(\mathbf{a})}{P_0(\mathbf{a})}\mu_{2m}^{(2)}n_2 + \cdot + \frac{B_0^l(\mathbf{a})}{P_0(\mathbf{a})}\mu_{2m}^{(l)}n_l + \cdot \\
&\quad + \frac{B_0^{(2m+1)}(\mathbf{a})}{P_0(\mathbf{a})}\mu_{2m}^{(2m+1)}n_{2m+1} = H^{2m} + K. \qquad (101)
\end{aligned}
$$

Let $\mathbf{M}$ denote the $(2m+1)\times(2m+1)$ matrix whose generic term is $[B_0^l(\mathbf{a})/P_0(\mathbf{a})]\mu_k^{(l)}$, $k = 0, \ldots, 2m$; $l = 1, \ldots, 2m + 1$. Each of those elements is the kth moment of bond l weighted by the ratio of bond l's initial value $B_0^l(\mathbf{a})$ to the portfolio's initial value $P_0(\mathbf{a})$. Let $\mathbf{n}$ be the $(2m + 1)$-dimensional column vector $(n_1, \ldots, n_l, \ldots, n_{2m+1})$. Finally, let $\mathbf{h}_{2m+1}^p$ stand for the column vector $(1, H, \ldots, H^k, \ldots, H^{2m} + K)$.

This is not the only way to envision system (101). An alternative is to consider the shares $n_l B_0^l/P_0 \equiv \alpha_l$ ($l = 1, \ldots, 2m + 1$) and then the square matrix of $(2m + 1)^2$ elements $\alpha_l \mu_k$ ($l = 1, \ldots, 2m + 1$; $k = 0, \ldots, 2m$), each of those elements corresponding to line k and column l respectively. This matrix multiplies the column vector of the α_l's, the unknowns. The system is thus solved in terms of the shares α_l instead of the numbers n_l. Both methods yield equivalent results.

System (101) becomes simply

$$\mathbf{Mn} = \mathbf{h}_{2m+1}^p \qquad (102)$$

which can be solved in n if and only if $\mathbf{M}$ is not singular. We get

$$\mathbf{n} = \mathbf{M}^{-1}\mathbf{h}^{p}_{2m+1}. \tag{103}$$

The Hessian matrix equals

$$P_{0,H}\boldsymbol{\Gamma} = P_{0,H}\begin{bmatrix} 0 & \cdots & & 0 \\ \vdots & \ddots & & \vdots \\ & & 0 & 0 \\ 0 & \cdots & 0 & K \end{bmatrix} = \begin{bmatrix} 0 & \cdots & & 0 \\ \vdots & \ddots & & \vdots \\ & & 0 & 0 \\ 0 & \ldots & 0 & P_{0,H}K \end{bmatrix} \tag{104}$$

and the associated quadratic form is

$$\mathrm{d}\mathbf{A}^{T}\begin{bmatrix} 0 & \cdots & & 0 \\ \vdots & \ddots & & \vdots \\ & & 0 & 0 \\ 0 & \cdots & 0 & P_{0,H}K \end{bmatrix}\mathrm{d}\mathbf{A} = P_{0,H}K\,\mathrm{d}A_m^2. \tag{105}$$

Taking into account the fact that the gradient of P_H at $\mathbf{a}$, $\nabla P_H(\mathbf{a})$, equals zero, the Taylor development of $P_H(\mathbf{A})$ around point $\mathbf{a}$ yields

$$P_H(\mathbf{A}) - P_H(\mathbf{a}) = P_{0,H}(\mathbf{a})K\,\mathrm{d}A_m^2 + R_2(\mathbf{a}, \mathrm{d}\mathbf{A}) \tag{106}$$

where $R_2(\mathbf{a}, \mathrm{d}\mathbf{A})$ denotes the remainder of the Taylor series. Since by definition $R_2(\mathbf{a}, \mathrm{d}\mathbf{A})$ is finite, we can always determine a value of K such that $P_H(\mathbf{A}) > P_H(\mathbf{a})$, thus making $P_H(\mathbf{A})$ convex at point $\mathbf{a}$ whatever the magnitude of the displacement of $\mathbf{A}$, $\mathrm{d}\mathbf{A}$. In applying our theorem, we will illustrate how increasing K increases only the *local* convexity of the portfolio's future value.

We will now show that it is impossible to achieve immunization simply by meeting first-order conditions. In other words, if one builds up a bond portfolio that conforms to first-order conditions only (all moments of order k, $k = 0, \ldots, m$, are equal to the kth power of the horizon), it will never be possible to achieve second-order conditions for a minimum of the portfolio's future value. The likelihood that a portfolio formed in such a way that $\gamma_0, \gamma_1, \ldots, \gamma_m = 0$ (such that it meets the first-order conditions) *and* that, in addition, it meets independently *any* of the conditions $\gamma_{m+1}, \gamma_{m+2}, \ldots, \gamma_{2m-1} = 0$, is nil. So is the probability that it meets all of them. Suppose then that one value of γ_k $(k = m+1, \ldots, 2m-1)$ differs ever so slightly from zero. Call this generalized variance γ. From what we have shown above, at least one line of elements parallel to the second diagonal of the Hessian matrix is made of γ's. There are two possibilities: either γ is on the first diagonal, or it is not. We take them up in order:

(a) γ is on the first diagonal. In that case γ must be positive. Also, there is a principal minor such as

$$\begin{vmatrix} 0 & 0 & \gamma \\ 0 & \gamma & \hat{\gamma} \\ \gamma & \hat{\gamma} & \gamma^{*} \end{vmatrix}$$

where $\hat{\gamma}$ and γ^* are as yet unspecified. The minor is equal to $-\gamma^3 < 0$; thus the portfolio cannot reach a minimum at **a**.

(b) γ is not on the first diagonal. Then there is a principal minor such as

$$\begin{vmatrix} 0 & \gamma \\ \gamma & \hat{\gamma} \end{vmatrix}$$

which is equal to $-\gamma^2 < 0$ whatever the sign of γ, a contradiction with necessary conditions for a minimum.

3.8. The nature of the cash flows of an immunizing portfolio

Equations (101) look innocuous enough. However, they carry deep implications about the nature of the cash flows of the immunizing portfolio. Indeed, they imply that if an immunizing portfolio exists, *it must have one or more negative cash flows*. This result was first shown, in discrete analysis, by Anthony Pakes (in a private communication to the author); for a proof in continuous analysis (supposing that bonds pays a continuous coupon), see Pakes (2001). Consider the second and the third equations of system (101). They stem from the first and the second equations of system (85), which can be written as

$$\sum_{t=1}^{N} t c_t \exp(-\mathbf{a} \cdot \mathbf{t}^p)/P(\mathbf{a}) = H \tag{85-1}$$

and

$$\sum_{t=1}^{N} t^2 c_t \exp(-\mathbf{a} \cdot \mathbf{t}^p)/P(\mathbf{a}) = H^2. \tag{85-2}$$

Both equations imply

$$\sum_{t=1}^{N} (t - H)^2 c_t \exp(-\mathbf{a} \cdot \mathbf{t}^p)/P(\mathbf{a}) = 0. \tag{85-3}$$

If Equation (85-3) holds, as well as system (101), at least one cash flow c_t must be negative. This is possible if and only if at least one position n_l $(l = 1, \ldots, L)$ is negative. We will be able to verify, in the following applications, that immunizing portfolios *always* carry some negative cash flows, and therefore that immunizing portfolios have always at least one negative position.

4. Applications

Suppose we observe today a spot structure which, for reasons explained in Section 1, is represented by a third-order polynomial. The future spot rate curve, i.e., the structure

that will prevail in, say, one year, is unknown and is assumed to remain unpredictable. We now show with concrete examples how an immunizing portfolio is to be constructed, and we will put this portfolio to the test of very strong variations of the initial spot rate structure. In fact, those variations will even amount, in some cases, to changing the structure from a steeply increasing one to a decreasing one, and vice versa.

4.1. The spot structures and their shifts

We consider eight arbitrary spot structures, labeled as $A, B, C, \ldots, G, H$ (this last designation is independent from that of the immunization horizon H); see Table 1. To each spot rate structure corresponds a triplet of points through which the spot rate curve passes.

Each of those curves is supposed to level off for maturity $t = \bar{s} = 20$ years. In order to find the coefficients of the corresponding polynomials, we have to solve system of Equations (38). Doing so yields the following polynomials:

$$A\colon s_A(t) = 1.1\big(10^{-6}\big)t^3 - 9.4\big(10^{-5}\big)t^2 + 0.002444t + 0.04, \tag{107}$$

$$B\colon s_B(t) = 0.06,$$

$$C\colon s_C(t) = 6.7\big(10^{-7}\big)t^3 - 5.7\big(10^{-5}\big)t^2 + 0.001467t + 0.048, \tag{108}$$

$$D\colon s_D(t) = 3.9\big(10^{-7}\big)t^3 - 5.3\big(10^{-5}\big)t^2 + 0.001656t + 0.04, \tag{109}$$

$$E\colon s_E(t) = 1.2\big(10^{-6}\big)t^3 - 7.9\big(10^{-5}\big)t^2 + 0.001767t + 0.053, \tag{110}$$

$$F\colon s_F(t) = -3.9\big(10^{-7}\big)t^3 - 2.7\big(10^{-5}\big)t^2 + 0.001544t + 0.035, \tag{111}$$

$$G\colon s_G(t) = -8.3\big(10^{-7}\big)t^3 + 0.000051t^2 - 0.00103t + 0.052, \tag{112}$$

$$H\colon s_H(t) = -7.2\big(10^{-7}\big)t^3 + 1.4\big(10^{-6}\big)t^2 + 0.000811t + 0.049. \tag{113}$$

Table 1
Basic spot rate structures, as defined by series of points $(t, s(t))$

Maturity t (years)	Spot rates $s(t)$ for various structures (per year, continuously compounded)			
	$s_A(t)$	$s_B(t)$	$s_C(t)$	$s_D(t)$
0	0.04	0.06	0.048	0.04
5	0.05	0.06	0.054	0.047
$t = \bar{s} = 20$	0.06	0.06	0.06	0.055
Maturity t (years)	Spot rates $s(t)$ for various structures (per year, continuously compounded)			
	$s_E(t)$	$s_F(t)$	$s_G(t)$	$s_H(t)$
0	0.053	0.035	0.052	0.049
5	0.060	0.042	0.048	0.053
$t = \bar{s} = 20$	0.066	0.052	0.045	0.06

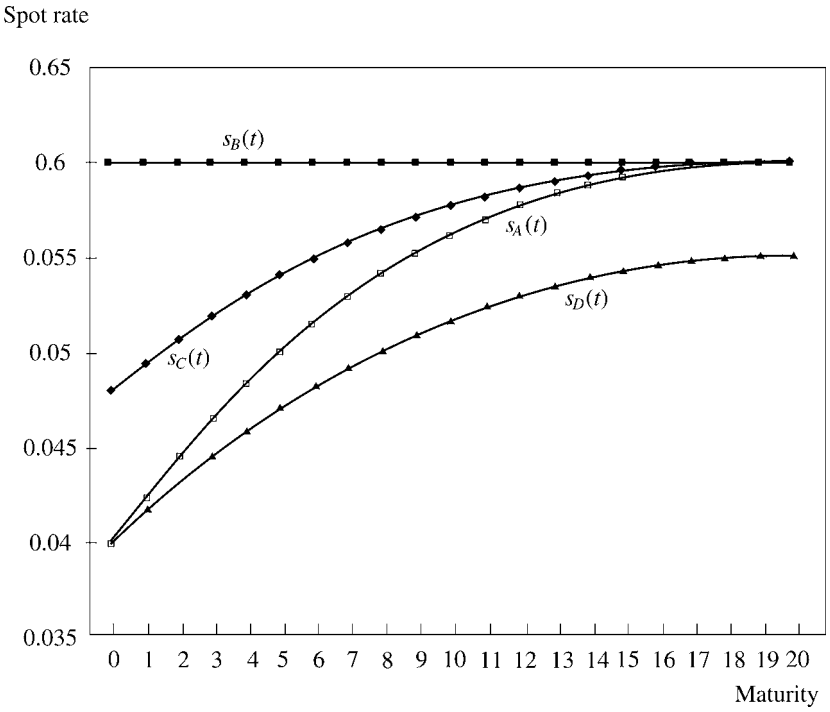


Fig. 7. Spot structures $s_A(t)$, $s_B(t)$, $s_C(t)$ and $s_D(t)$ (continuously compounded yearly rates).

These eight structures are represented in Figures 7 and 8. They can be combined into $8 \times 7 = 56$ pairs: $AB, AC, \ldots, HG$ each pair designating a scenario. The first letter corresponds to the initial structure; the second one designates the structure that is observed immediately *after* the bond portfolio's purchase. Any of those 56 scenarios can generate an infinity of sub-scenarios, in the following way.

Let $\lambda \in [-1, 2]$ denote a parameter (the definition interval of λ is arbitrary; it could be made larger or smaller). Let $s_a(t)$ be an initial spot rate structure; let $s_b(t)$ designate a new one. Consider the spot rate defined by

$$s_{ab,\lambda}(t) = s_a(t) + \lambda\big[s_b(t) - s_a(t)\big] = (1 - \lambda)s_a(t) + \lambda s_b(t). \tag{114}$$

Thus $s_{ab,\lambda}(t)$ is a linear combination of the spot rates s_a and s_b; particular cases correspond to $\lambda = 0$ ($s_{ab,0}(t) = s_a(t)$, the initial structure) and $\lambda = 1$ ($s_{ab,1}(t) = s_b(t)$, the "new" structure). The initial spot structure s_a and the new spot rate s_b, together with λ, define a direction of change in the space of the polynomials' coefficients. This directional vector is simply $\lambda(\mathbf{A} - \mathbf{a})$ where $\mathbf{a}$ is the vector of the coefficients of the initial polynomial ($s_A(t)$ in this case) and $\mathbf{A}$ is the vector of the coefficients of the new polynomial ($s_B(t)$ here).

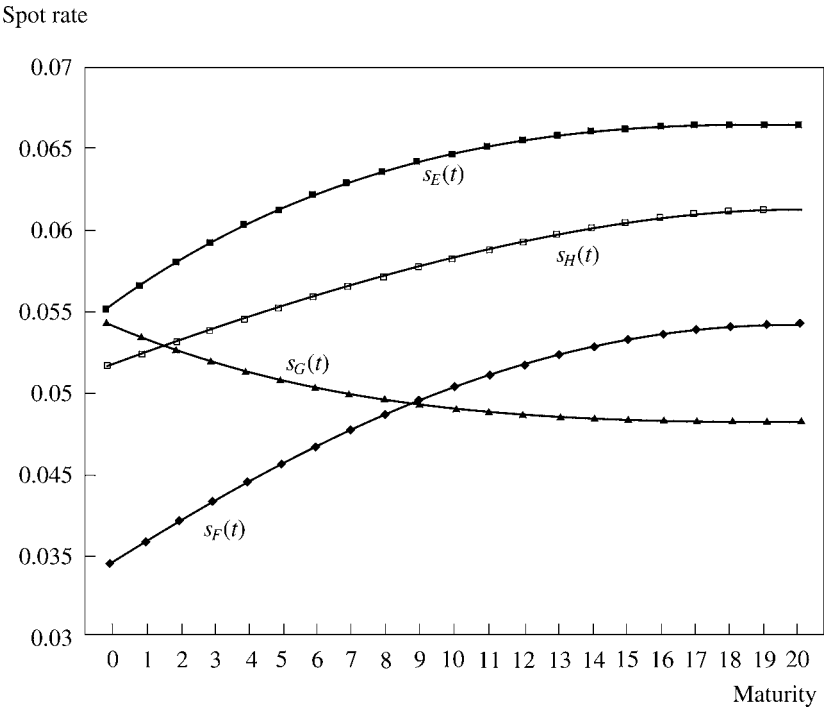


Fig. 8. Spot structures $s_E(t)$, $s_F(t)$, $s_G(t)$ and $s_H(t)$ (continuously compounded yearly rates).

An infinity of variations of the spot rate curve can thus be generated by changing λ. Figure 9 illustrates such variations with λ taking the values -1; -0.5; 0; 0.5; 1; 1,5. The initial spot rate curve ($\lambda = 0$) is s_D: it is steeply increasing (from 4% for short rates to 5.5% for long ones). The new structure ($\lambda = 1$) is s_B, a horizontal at 6%. The values $\lambda = -1$ and $\lambda = -0.5$ correspond to a decrease of all rates; the value $\lambda = 0.5$ yields intermediate spot rates between s_D and s_B. As to $\lambda = 1.5$, it leads to an inversion of the structure, i.e., to a decreasing spot rate. Indeed, from the definition

$$s_{ab,\lambda}(t) = s_a(t) + \lambda\big[s_b(t) - s_a(t)\big]$$

we have

$$s'_{ab,\lambda}(t) = s'_a(t) + \lambda\big[s'_b(t) - s'_a(t)\big],$$

from which we deduce immediately the rule about the slope of $s_{ab,\lambda}(t)$:

- if $s'_b > s'_a$

$$s'_{ab,\lambda}(t) \gtreqless 0 \quad \text{if and only if} \quad \lambda \gtreqless \frac{s'_a(t)}{s'_a(t) - s'_b(t)};$$

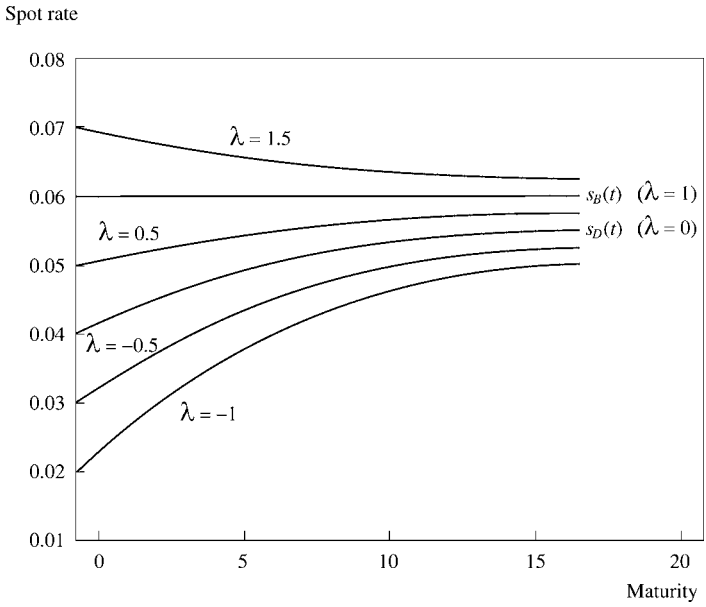


Fig. 9. Varying parameter λ generates an infinity of spot rate curves.

- if $s'_b < s'_a$

$$s'_{ab,\lambda}(t) \gtreqless 0 \quad \text{if and only if} \quad \lambda \lesseqgtr \frac{s'_a(t)}{s'_a(t) - s'_b(t)};$$

- if $s'_b = s'_a$, $s'_{ab,\lambda}(t) = s'_a(t) = s'_b(t)$.

In our example, $s'_a(t) = s'_D(t) > 0$ and $s'_b(t) = s'_B(t) = 0$; so we are in the case $s'_b < s'_a$, and we have an inversion of the spot rate curve if and only if $\lambda > 1$.

We are now ready to describe the immunization process by which our investor can protect himself against *any* change in the structure of spot rates.

4.2. Building immunizing portfolios

Suppose that we have at our disposal 9 bonds, labeled B_l, $l = 1, \ldots, 9$, whose characteristics are summarized in Table 2.

Suppose also that our initial spot rate curve is

$$s_A(t) = 1.1\big(10^{-6}\big)t^3 - 9.4\big(10^{-5}\big)t^2 + 0.002444t + 0.04.$$

We have to determine the vector $\mathbf{n} = n_1, \ldots, n_9$ that solves system

$$\mathbf{Mn} = \mathbf{h}^p.$$

Table 2
Basic characteristics of bonds used in immunization portfolio

Bond number	1	2	3	4	5	6	7	8	9
Coupon rate (in percent)	4	4.75	7	8	6	4.5	7	5	5.5
Maturity (years)	7	8	15	20	6	9	16	12	10

Then

$$\mathbf{n} = \mathbf{M}^{-1}\mathbf{h}^{p}.$$

Our first task is to build matrix $\mathbf{M}$ whose elements are $[B_0^l(a)/p_0(a)]\mu_k^{(l)}$, $k = 0, \ldots, 2m$; $l = 1, \ldots, 2m+1$. Using the spot structure $s_A(t)$, we obtain the values of the bonds and their various moments. The results are presented in Table 3, which can be viewed as a (9×9) matrix. Recall that the moment of order k is expressed in $(\text{years})^k$.

The (1×9) first line of matrix $\mathbf{M}$ is the first line of Table 3 (the initial value of each bond expressed in terms of its par value); each subsequent line of matrix $\mathbf{M}$ is the corresponding line of Table 3, scaled by the value of each bond in par value. Matrix $\mathbf{M}$ can thus be seen as Table 3 premultiplied by the diagonal matrix of the values of the bonds.

Finally, suppose our immunization horizon is 7 years. We know that all generalized variances of order 0 to $2m-1$ should be equal to zero, and that the generalized variance of order $2m$, γ_{2m}, should be positive. Since $\gamma_{2m} = \gamma_8 = \mu_8 - h^8$, the way to achieve this is to equate μ_8 to H^8 plus a positive constant K. Set this constant arbitrarily to 200,000 (expressed in $(\text{years})^8$). This amounts to setting K at a relatively low level, equal to about 3.4% of H^8. Further we discuss the consequences, for the immunizing process, of setting K at various levels. For the time being, we can rest assured that a local minimum will be attained for the portfolio's future value. Inverting matrix $\mathbf{M}$, $\mathbf{n}$ gives

$$\mathbf{n} = \mathbf{M}^{-1} \cdot \mathbf{h}$$

where

$$\mathbf{h} = \begin{bmatrix} 1 \\ 7 \\ 49 \\ 343 \\ 2{,}401 \\ 16{,}807 \\ 117{,}649 \\ 823{,}543 \\ 5{,}964{,}801 \end{bmatrix}.$$

Table 3
Value and moments of order $k = 1, \ldots, 2m = 8$ of each bond under spot rate structure: $s_A(t) = 1.1(10^{-6})t^3 - 9.4(10^{-5})t^2 + 0.002444t + 0.04$

Bond	B_1	B_2	B_3	B_4	B_5
Initial value[a]	0.9215	0.9537	1.1085	1.2343	1.0390
Moments					
μ_1	6.1940	6.7878	9.8975	11.3019	5.2211
μ_2	41.300	50.853	124.486	175.241	29.569
μ_3	282.14	393.60	1703.4	3058.9	172.29
μ_4	1946.9	3087.8	24174	56300	1016.0
μ_5	13501	24388	349522	1.065209	6028
μ_6	93,898	193,361	5,108,123	20,479,893	35,884
μ_7	654,177	1,536,761	75,155,699	$4.0(10^8)$	214,086
μ_8	4,562,925	12,232,997	$1.1(10^9)$	$7.8(10^9)$	1,279,087
Bond	B_6	B_7	B_8	B_9	
Initial value	0.9250	1.1099	0.9505	0.9755	
Moments					
μ_1	7.5036	10.2761	8.0123	8.9826	
μ_2	62.849	136.15	73.349	95.892	
μ_3	545.57	1973.91	701.47	1084.2	
μ_4	4806.3	29757	6831.2	12556	
μ_5	42655	457680	67134	147197	
μ_6	380,145	7,120,937	663,180	1,737,533	
μ_7	3,396,730	$1.1(10^8)$	6,572,335	20,598,422	
μ_8	30,403,315	$1.8(10^9)$	65,273,038	$2.4(10^8)$	

[a]This initial value is expressed in terms of its par value B_T. It is equal to B_0/B_T. Moments of order k are expressed in (years^k).

Performing the necessary calculations yields the immunizing portfolio:

$$n_1 = 6.161, \quad n_4 = 0.009, \quad n_7 = -0.448,$$
$$n_2 = -6.534, \quad n_5 = -1.211, \quad n_8 = 1.512,$$
$$n_3 = 0.823, \quad n_6 = 2.474, \quad n_9 = -1.371.$$

4.3. Immunization results

We now have to check that this portfolio protects us against *any* change in the spot rate structure. To that effect we will consider all scenarios $AB, AC, \ldots, AH$, and for each scenario a set of λ values from -1 to $+1.5$. Table 4 provides the results. The first figure in each cell indicates the portfolio's value just after the change from the initial structure to the new one; the second figure is the future portfolio's value based upon the new structure. Consider, for instance, the third cell of the first line, under $\lambda = 0.5$. It carries

Table 4
Immunization results for initial spot rate $s_A(t)$. Present value of portfolio at time 0: 100. Future value protected at time 0: 144.7809;[a] $K = 200{,}000$

Scenario	$\lambda = -1$	$\lambda = -0.5$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 1.5$
AB	105.1217	102.5289	97.5335	95.1278	92.7815
	144.7809	144.7809	144.7809	144.7809	144.7809
AC	102.0181	101.044	99.006	98.0219	97.0475
	144.7809	144.7809	144.7809	144.7809	144.7809
AD	97.4147	98.6989	101.3183	102.6539	104.0072
	144.7809	144.7809	144.7809	144.7809	144.7809
AE	106.52	103.2085	96.8912	93.8791	90.9606
	144.7809	144.7809	144.7809	144.7809	144.7809
AF	94.2190	97.0664	103.0222	106.1358	109.3435
	144.7809	144.7809	144.7809	144.7809	144.7809
AG	95.9591	97.0664	102.0839	104.2112	106.3829
	144.7809	144.7809	144.7809	144.7809	144.7809
AH	101.1502	100.5734	99.4299	98.8630	98.2994
	144.7809	144.7809	144.7809	144.7809	144.7809

First figure in each cell: present value of portfolio at time ε, under new spot structure.
Second figure: future value of portfolio under at time ε, under new spot structure.
[a]This value is equal to $100\exp[s_A(7)\cdot 7]$.

two numbers: the first figure (97.5335) indicates the new present value of the portfolio at time ε, i.e., for a structure half-way between structures A and B, and thus equal to

$$s_{AB,0.5}(t) = s_A(t) + 0.5\big[s_B(t) - s_A(t)\big].$$

The second figure (144.7809) is the future value of the portfolio at horizon 7 years under the new structure $s_{AB,0.5}(t)$: it tells us that the portfolio is perfectly immunized since that is the future value of the portfolio before the spot structure change.

From the results in Table 4, we conclude that immunization is obtained in each of the scenarios starting with spot rate structure $s_A(t)$. The same conclusion would be reached should the initial spot structure be $s_B(t), \ldots, s_H(t)$. Table 5 indicates, for each of those initial structures, the future value protected, as well as the corresponding portfolio.

A remarkable, far from intuitive, conclusion emerges: the optimal portfolios are *little sensitive* to different initial (observed) structures. Even quite different initial spot curves such as $s_A(t)$, $s_E(t)$, or $s_G(t)$ are not conducive to substantial modifications in the immunizing portfolios. Their orders of magnitude remain the same (within boundaries of 3%), and, of course, the position signs remain the same.

This weak sensitivity of the portfolio to the initial spot structure implies that the cash flows generated by an asset are more important than the exact discounting factors in the evaluation process: in other words, what counts most is the cash flow profile of each bond. We will confirm this when we look for the optimal portfolio that would be built up from different sets of bonds. We will introduce variations first in the coupons, then

Table 5
Future value protected and immunizing portfolio for each initial spot rate structure $s_A(t), \ldots, s_H(t)$; Horizon $H = 7$ years; $K = 200{,}000$

Initial structure	Future value protected	Immunizing portfolio				
		n_1	n_2	n_3	n_4	
		n_5	n_6	n_7	n_8	n_9
$s_A(t)$	144.7809	6.161	−6.534	0.823	0.009	
		−1.211	2.474	−0.448	1.512	−1.371
$s_B(t)$	152.1962	6.475	−6.857	0.833	0.009	
		−1.273	2.598	−0.452	1.562	−1.407
$s_C(t)$	147.7026	6.285	−6.662	0.827	0.009	
		−1.236	2.523	−0.450	1.532	−1.386
$s_D(t)$	141.0378	5.993	−6.306	0.765	0.008	
		−1.186	2.355	−0.414	1.460	−1.297
$s_E(t)$	154.2206	6.5724	−7.0296	0.9055	0.0098	
		−1.2828	2.7049	−0.4953	1.6101	−1.4893
$s_F(t)$	136.411	5.7910	−6.0620	0.7273	0.0074	
		−1.1510	2.2419	−0.3935	1.4103	−1.2402
$s_G(t)$	138.9303	5.8877	−6.0986	0.6679	0.0064	
		−1.1782	2.2214	−0.3566	1.3961	−1.1890
$s_H(t)$	146.446	6.2298	−6.5922	0.8216	0.0087	
		−1.2272	2.4904	−0.4476	1.5173	−1.3713

in maturities, and finally in both parameters. Prior to that, we discuss the question of the magnitude of the immunization parameter K.

4.4. How large should we set the immunization parameter K?

We introduced parameter K to set the generalized variance $\gamma_{2m+1} = \gamma$ at a positive number. We had set arbitrarily K at 200,000. For that value of K, it is always possible to protect the investor against any change generated in any possible scenario $AB, \ldots, HG$ for values of λ between -1 and $+1.5$. This means that immunization is warranted *at least* in this interval. We will now study this relationship.

We have to keep in mind that the immunization process has the happy effect of making the future value $F_H(\lambda)$ an extremely flat curve in F_H, λ space, and we have to ask the question whether the fact that the portfolio's future value is guaranteed within 4 decimal places (as Table 4 reveals) warrants that this future value is *strictly* higher than the value to be protected. This prods us to examine each figure with more than 4 decimals. The conclusion is that for all scenarios these future values are indeed strictly larger than the aim, with one exception: if the initial structure is D and counter-pivots with respect to structure B, this strict inequality does not hold for values of λ such that $\lambda < -0.346$—see Table 6. Of course, practitioners would be happy to protect an investment as well as in case $\lambda = -1$, for instance: Table 6 tells us that the future value of an initial $100,000,000 portfolio is immunized against a very large, non-parallel shift

Table 6
Results of immunization; scenario *DB*; horizon $H = 7$ years; $K = 200{,}000$; Target: $B_{0,H} = 141{,}037{,}819.08 = 100\exp[7s_D(7)]$

λ	$B_H\{s_D(t) + \lambda[s_B(t) - s_D(\lambda)]\}$	
1	141,037,827.60	values higher or equal to target
0.8	141,037,823.70	
0.6	141,037,821.22	
0.4	141,037,819.83	
0.2	141,037,819.22	
0	141,037,819.08	
−0.2	141,037,819.12	
−0.346[a]	141,037,819.03	
−0.4	141,037,819.03	values lower than target
−0.6	141,037,818.53	
−0.8	141,037,817.33	
−1	141,037,815.15	

$\lambda = -0.34596239916$.
[a]Rounded value.

of the spot curve within a margin smaller than \$4. Nevertheless, we want to study this case on the following grounds: it is important to be able to picture $B_H(\lambda)$ as a *locally convex* curve of λ, with a local minimum at $\lambda = 0$, and to see how this curve is modified when changing K. From what we have shown previously, K should play the role of a local convexity parameter: the higher K, the more convex should be the curve in F_H, λ space, and our aim now is to verify this.

We shall first consider a close-up of the situation within a close vicinity of $\lambda = 0$, i.e., in the interval $\lambda \in [-0.2, +0.2]$, with the following immunization constants: $K = 200{,}000$; $K = 100{,}000$ and $K = 50{,}000$. The results conform exactly our predictions: a higher value of K enhances convexity of the $P_H(\lambda)$ curve in a vicinity of the minimum point $(0, P_H(0))$ as evidenced in Figure 9.

Now let us pay attention to this family of curves. Observe that for positive values of λ, the curves exhibit strong convexity, whilst for negative values convexity is somewhat weaker, transforming relatively quickly into concavity. This leads us to surmise that, if complete immunization is easily achieved for positive values of λ, the portfolio's future value may well decline and even become slightly smaller than the target if λ is sufficiently large in absolute value. This indeed turns out to be the case as evidenced by Figure 12 where we have considered values of λ from -0.6 to $0.2\dot{4}$ If λ is sufficiently small, the future value starts declining and ultimately becomes lower – if ever slightly than $F_H(\mathbf{a})$. Let us remember, however, the orders of magnitude: in our example, the future value (at horizon 7 years) of \$100,000,000 today remains immunized within \$4! One striking feature of Figures 10–12 is to throw light onto the fact that K

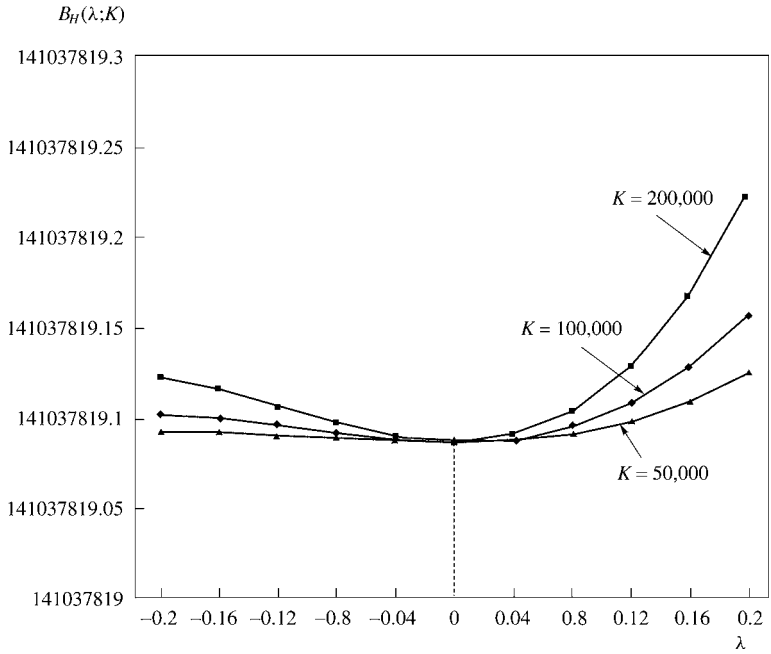


Fig. 10. Increasing K enchances local convexity of the portfolio's future value.

is a local convexity parameter: its sole predictable effect is to increase convexity at the initial point **a** (equivalently: for $\lambda = 0$). (Observe how, around $\lambda = -0.36$, the curves intersect.)

We conclude from this analysis, and the one that can be carried out from a wider range of initial spot structures, that it is enough to use an immunization constant K in the order of 3% of the last moment.

4.5. Infinity of solutions

Considering the basic system of $2m + 1$ Equations (101) from which an immunizing portfolio is determined, it is obvious that if L, the number of bonds available, is larger than $2m + 1$ there is an infinity of solutions.

Pakes (2001) has a deep discussion of these issues. In this section, we consider that either $L = 2m + 1$, or, if $L > 2m + 1$, the L bonds are arranged in $2m + 1$ sub-groups of bonds. We discuss in this section the infinity of solutions resulting from the fact that the convexity constant $K\,(K > 0)$ is arbitrary.

We are now ready to state our last theorem, concerning the number of solutions that correspond to the various possible values of K.

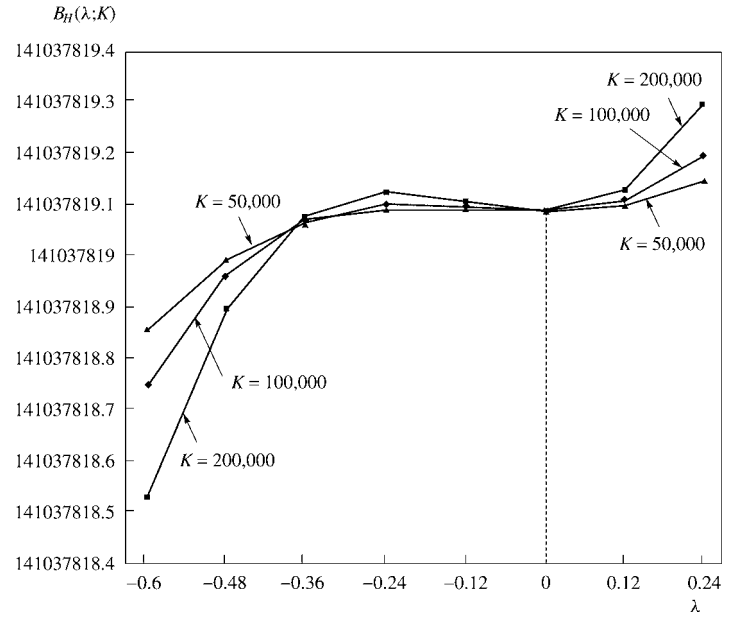


Fig. 11. While increasing local convexity, higher values of K do not necessarily widen immunization interval.

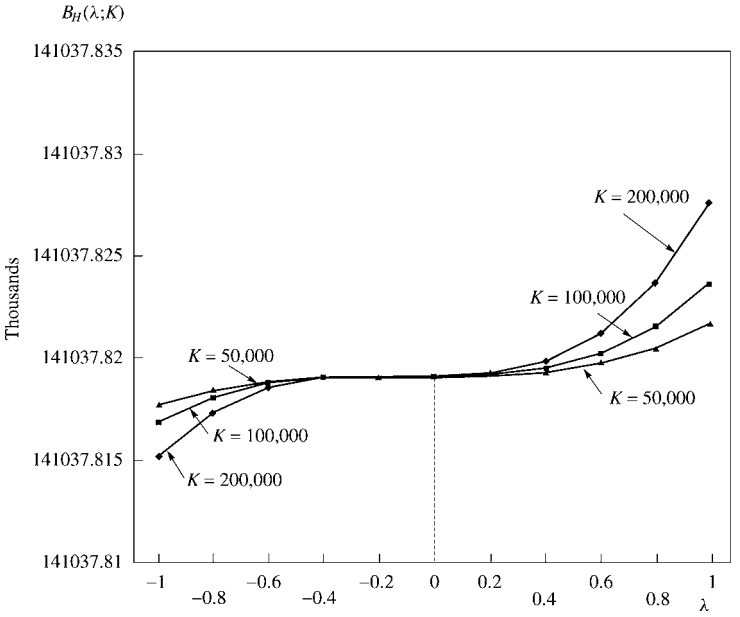


Fig. 12. At a wider scale, F_H may look going through an inflection point at $\lambda = 0$, although it undergoes a local minimum there.

Theorem 6. *Suppose that the spot rate structure can be expanded into a Taylor series. Then there exists an infinity of immunizing bond portfolios.*

The proof of this theorem rests upon the continuity of the immunization constant K.

The fact that an infinity of bond portfolios can protect the investor can have important practical applications. As indicated before, the outcome of the immunization process is to make the future value of the portfolio an extremely "flat" curve in (P_H, λ) space. Therefore the interest of having large choices for K will not be to try to enhance convexity; it will rather reside in the extended available choices when building portfolios.

A natural way to apply this important property could be to choose K such as to minimize the transaction costs of building up, or adjusting the bond portfolio.

4.6. How sensitive are immunizing portfolios to changes in horizon H?

We now address the all-important question of the sensitivity of immunizing portfolios to changes in the horizon. To answer that question, we have chosen to study optimal portfolios corresponding to the initial spot rate structures $s_A(t)$, $s_B(t)$ and $s_G(t)$, for horizons ranging from three years to twenty years. This span of years for horizon H could have been wider, because theory tells us that immunizing portfolios do exist for a broader range of horizons. However, since the basket of bonds to chose from is the initial one, with maturities between 6 and 20 years, we may surmise that overly large negative and positive positions would be the rule if we attempted to perform immunization for very long or very short horizons with bonds to be chosen from such a basket. We will be able to confirm this.

For the time being, let us examine the results we get when immunizing under the initial spot rate structure $s_A(t)$. The portfolios for each horizon are indicated in Table 7, and their evolution with respect to the horizon are pictured in Figure 13 (bonds 1–5) and in Figure 14 (bonds 6–9). It is quite striking to notice how stable the portfolios remain when the horizons are in the 6–16 years range. When horizons exceed that range, the portfolios become highly unstable.

Our initial structure $(s_A(t))$ was an increasing one. We can now confirm these results in the case of a constant structure $(s_B(t))$ and a decreasing one $(s_G(t))$ (see Tables 8 and 9, and the corresponding Figures 15–18). The similarity of the portfolios for any initial spot rate is striking. It generalizes the property we had shown previously for all spot structures from $s_A(t)$ to $s_G(t)$ when the horizon was 7 years. This property applies to all immunization horizons. Furthermore, for any kind of initial spot rate, portfolios are very stable for mid-range horizons, and become highly unstable outside that range.

Table 7
Immunizing portfolios according to horizon H, under initial spot rate structure $s_A(t)$; $K = 200{,}000$; number of each bonds in portfolio

Horizon H (years)	Immunizing portfolios								
	n_1 4% 7 years	n_2 4.75% 8 years	n_3 7% 15 years	n_4 8% 20 years	n_5 6% 6 years	n_6 4.5% 9 years	n_7 7% 10 years	n_8 5% 12 years	n_9 5.5% 10 years
3	−26.34	−26.98	−7.60	−0.05	15.93	103.35	3.65	−82.53	22.29
4	−8.69	−28.47	−3.47	−0.02	9.46	67.76	1.62	−47.72	11.38
5	2.45	−18.19	−0.40	0.001	3.24	26.81	0.13	−14.96	2.50
6	6.25	−9.92	0.71	0.008	−0.11	6.52	−0.40	−0.67	−0.96
7	6.16	−6.53	0.82	0.009	−1.21	2.47	−0.45	1.51	−1.37
8	5.36	−6.56	0.77	0.008	−1.33	4.26	−0.43	0.55	−1.19
9	5.23	−7.80	0.78	0.008	−1.30	5.48	−0.43	0.74	−1.22
10	5.70	−8.90	0.75	0.008	−1.39	5.13	−0.41	1.71	−1.09
11	6.13	−9.56	0.64	0.008	−1.49	5.04	−0.37	1.40	−0.30
12	6.21	−10.14	0.71	0.008	−1.48	6.64	−0.40	−0.95	0.91
13	6.22	−11.04	1.30	0.01	−1.43	9.27	−0.60	−3.79	1.56
14	6.69	−12.24	2.32	0.01	−1.51	10.76	−0.80	−4.52	0.75
15	7.57	−13.20	2.71	0.01	−1.74	10.03	−0.30	−2.64	−1.05
16	7.55	−13.15	0.29	0.01	−1.73	9.97	2.12	−2.61	−1.06
17	4.24	−11.75	−7.22	0.06	−0.72	17.87	7.47	−13.41	5.03
18	−3.50	−9.80	−19.06	0.3	1.76	39.06	14.73	−40.30	18.80
19	−9.58	−9.73	−25.21	1.04	3.84	58.68	17.15	−63.20	29.24
20	10.74	−15.28	4.50	2.97	−2.67	4.70	−3.83	5.35	−5.52

4.7. How sensitive are immunizing portfolios to a change in the basket of available bonds?

We now show that the slightest modification in the composition of the basket of available bonds has major consequences on the immunizing portfolio.

We revert to our base case: immunization with an initial spot rate $s_A(t)$; a horizon $H = 7$ years and a constant $K = 200{,}000$. We consider an innocuous-looking, very small change in the features of two among the nine bonds that were available for our initial application, as described in Section 4.2, Table 2. The last seven bonds are identical in their maturity and coupon. The maturities of bonds 1 and 2 are identical (7 and 8 years, respectively, and let us simply interchange their coupons: the coupon of bond 1 becomes 4.75 and the coupon of bond 2 becomes 4). The results are quite surprising: as Table 10 reveals, the immunizing portfolio undergoes huge shifts; for example n_8 increases from 1.5 to 26.3, and some positions change sign: bonds 1, 2, 5 and 6.

We can do the same exercises with bonds 6 and 9: interchanging their coupons while keeping the same maturities leads to the portfolio described in the third line of Table 10. (Notice that in the cases considered above, interchanging coupons is tantamount to exchanging maturities, as the reader can verify from Table 2, Section 4.2.) As to the

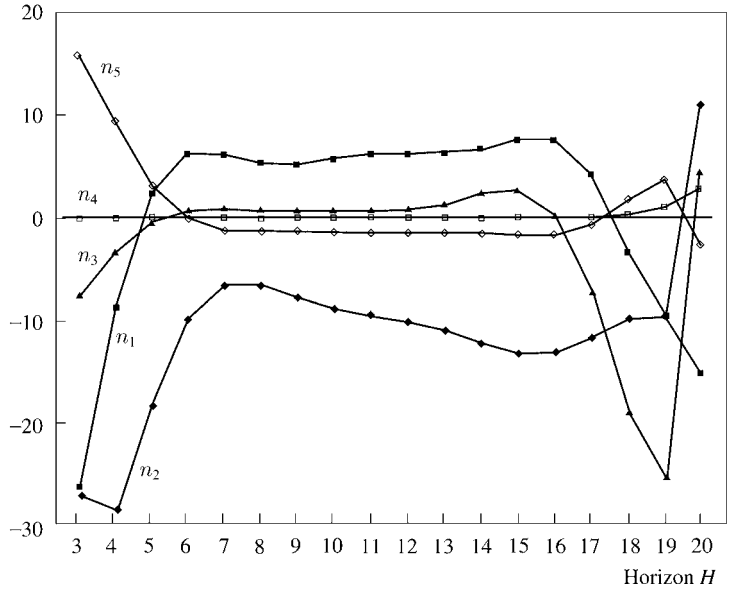


Fig. 13. Immunizing portfolios as a function of horizon H; initial spot structure: s_A; bonds 1–5.

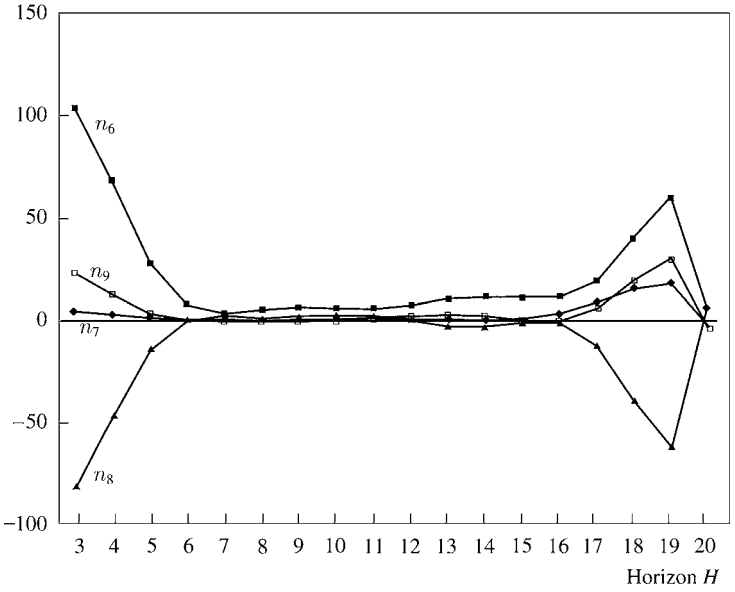


Fig. 14. Immunizing portfolios as a function of horizon H; initial spot structure: s_A; bonds 6–9.

Table 8
Immunizing portfolios according to horizon H, under initial spot rate structure $s_B(t)$; $K = 200{,}000$; number of each bonds in portfolio

Horizon H (years)	Immunizing portfolios								
	n_1 4% 7 years	n_2 4.75% 8 years	n_3 7% 15 years	n_4 8% 20 years	n_5 6% 6 years	n_6 4.5% 9 years	n_7 7% 10 years	n_8 5% 12 years	n_9 5.5% 10 years
3	−27.46	−29.21	−7.76	−0.05	16.78	109.35	3.71	−86.68	23.12
4	−9.03	−30.23	−3.54	−0.02	9.95	71.31	1.64	−49.93	11.77
5	2.62	−19.19	−0.41	0.001	3.40	28.17	0.17	−15.65	2.59
6	6.58	−10.43	0.72	0.008	−0.12	6.85	−0.40	−0.72	−0.98
7	6.48	−6.86	0.83	0.009	−1.27	2.60	−0.45	1.56	−1.41
8	5.62	−6.87	0.78	0.008	−1.40	4.45	−0.43	0.57	−1.22
9	5.47	−8.14	0.79	0.008	−1.37	5.68	−0.43	0.79	−1.26
10	5.94	−9.24	0.76	0.008	−1.45	5.25	−0.42	1.83	−1.13
11	6.37	−9.87	0.65	0.008	−1.55	5.08	−0.37	1.55	−0.33
12	6.43	−10.40	0.72	0.008	−1.54	6.67	−0.40	−0.85	0.92
13	6.41	−11.26	1.33	0.01	−1.49	9.32	−0.61	−3.75	1.57
14	6.85	−12.43	2.37	0.01	−1.56	10.77	−0.82	−4.45	0.72
15	7.74	−13.33	2.76	0.01	−1.79	9.89	−0.31	−2.43	−1.14
16	7.69	−13.23	0.31	0.01	−1.78	9.76	2.13	−2.36	−1.15
17	4.22	−11.79	−7.29	0.06	−0.71	18.04	7.53	−13.60	5.12
18	−3.85	−9.87	−19.28	0.29	1.88	40.31	14.84	−41.63	18.80
19	−10.20	−9.87	−25.51	1.04	4.06	60.89	17.29	−63.20	19.29
20	10.94	−15.22	4.58	2.97	−2.74	4.99	−3.86	6.04	−5.75

immunizing results, it is of no surprise that they are as good as those corresponding to the initial basket.

5. Conclusion and suggestions

This study brought its fair share of (nice) surprises. First, a portfolio immunized against *any* change of the spot structure carries *always* at least one negative position. Also surprising is the simplicity of the central result: if an investor is protecting herself from any shift of a spot curve represented by an $m-1$ order polynomial, she just has to build a portfolio such that its generalized variances of order 1 to $2m-1$ are equal to zero, and its generalized variance of order $2m$ is equal to a positive, arbitrary number—which leads to the next shocker: even with a limited pool of $2m+1$ available bonds, there exists an infinity of immunizing portfolios for one given, initial, spot rate curve.

Equally unforeseen is the fact that the structure of an immunizing portfolio depends so little on the initial, observed structure, or on the investor's objective (her horizon), and so much on the *nature* of the bonds used in the portfolio. Changing if ever so marginally the characteristics of one or two bonds in the basket of the available bonds may alter dramatically the composition of the whole bond portfolio.

Table 9
Immunizing portfolios according to horizon H, under initial spot rate structure $s_G(t)$; $K = 200{,}000$; number of each bonds in portfolio

Horizon H (years)	Immunizing portfolios								
	n_1 4% 7 years	n_2 4.75% 8 years	n_3 7% 15 years	n_4 8% 20 years	n_5 6% 6 years	n_6 4.5% 9 years	n_7 7% 10 years	n_8 5% 12 years	n_9 5.5% 10 years
3	−27.86	−29.91	−7.82	−0.05	17.06	111.40	3.73	−88.14	23.43
4	−9.16	−30.78	−3.56	−0.02	10.11	72.51	1.65	−50.71	11.92
5	2.66	−19.51	−0.42	0.001	3.46	28.62	0.14	−15.89	2.62
6	6.68	−10.6	0.72	0.008	−0.12	6.96	−0.40	−0.73	−0.99
7	6.58	−6.97	0.84	0.009	−1.30	2.64	−0.45	1.58	−1.42
8	5.71	−6.98	0.79	0.008	−1.42	4.52	−0.43	0.58	−1.23
9	5.56	−8.26	0.80	0.008	−1.39	5.76	−0.44	0.80	−1.28
10	6.03	−9.37	0.76	0.008	−1.47	5.32	−0.42	1.86	−1.14
11	6.46	−10.00	0.65	0.008	−1.57	5.13	−0.37	1.59	−0.33
12	6.51	−10.52	0.73	0.008	−1.56	6.72	−0.40	−0.83	0.92
13	6.48	−11.37	1.33	0.01	−1.50	9.38	−0.61	−3.75	1.58
14	6.92	−12.52	2.38	0.01	−1.57	10.81	−0.82	−4.44	0.72
15	7.80	−13.40	2.78	0.007	−1.80	9.87	−0.31	−2.38	−1.16
16	7.75	−13.27	0.32	0.007	−1.79	9.72	2.13	−2.30	−1.17
17	4.22	−11.81	−7.32	0.06	−0.71	18.11	7.55	−13.69	5.16
18	−3.98	−9.87	−19.38	0.3	1.92	40.75	14.90	−42.12	19.50
19	−10.42	−9.88	−25.64	1.04	4.13	61.65	17.36	−66.30	30.35
20	11.01	−15.21	4.61	2.97	−2.76	3.76	−3.88	6.28	−5.83

Table 10
The sensitivity of immunizing portfolios to changes in coupon rates

Initial bond basket changes[a]	Immunizing portfolios								
	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9
$c_1 = 4$; $c_2 = 4.75$ (Initial basket)	6.16 (Initial portfolio)	−6.53	0.82	0.009	−1.21	2.47	−0.45	1.51	−1.37
$c_1 = 4.75$	−10.33	32.12	2.13	0.016	1.77	−43.5	−1.05	26.26	−6.08
$c_2 = 4$; $c_6 = 5.5$; $c_9 = 4.5$	−1.55	11.57	1.43	0.01	0.19	−18.84	−0.73	13.01	−3.62

[a]This column indicates changes in coupons of some of the initial bonds.

These results suggest that the next step, for the practitioner, would be to minimize transaction costs by making use of the fact that, even for a well-defined, limited, set of $2m+1$ bonds, an infinity of immunizing portfolios are available. Minimizing transaction

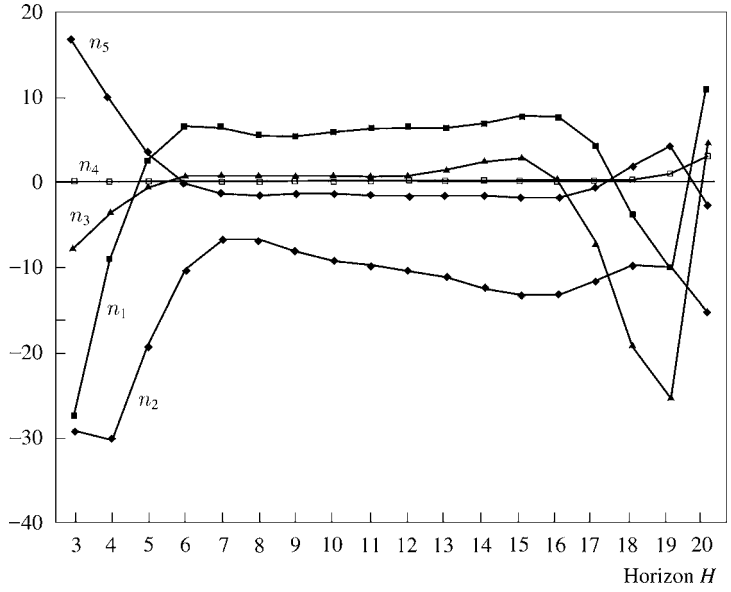


Fig. 15. Immunizing portfolios as a function of horizon H; initial spot structure: s_B; bonds 1–5.

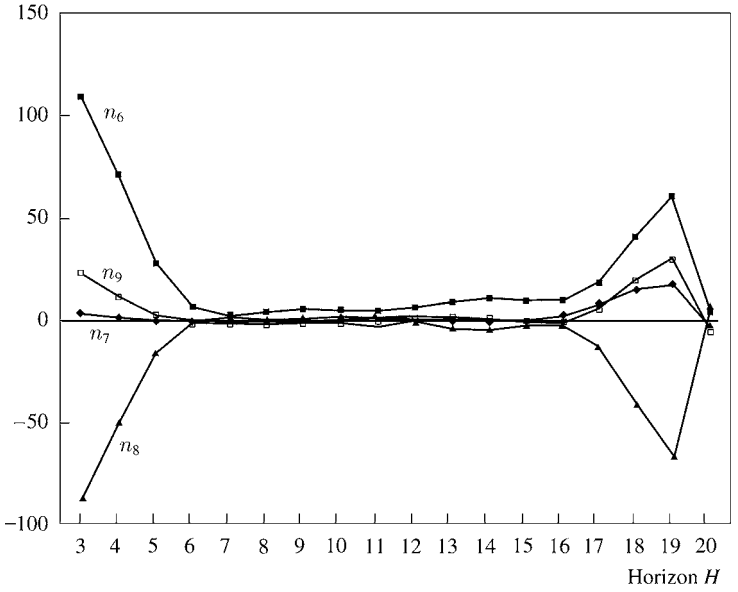


Fig. 16. Immunizing portfolios as a function of horizon H; initial spot structure: s_B; bonds 6–9.

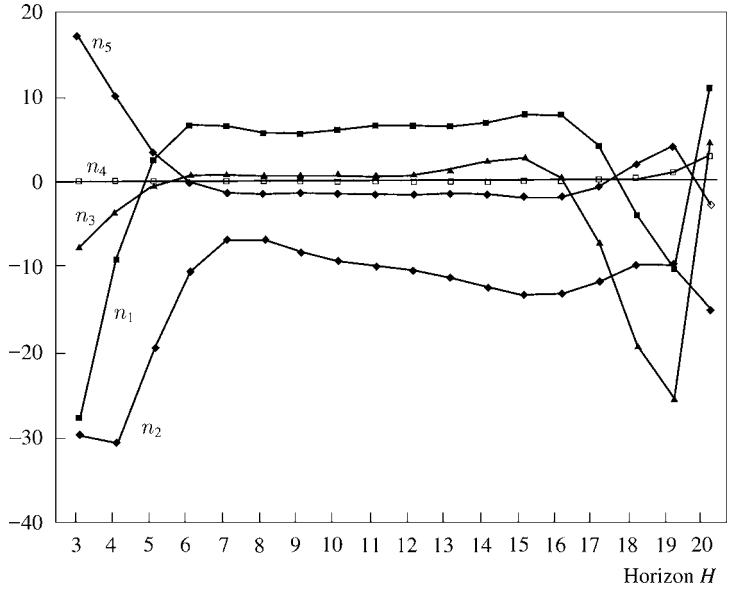


Fig. 17. Immunizing portfolios as a function of horizon H; initial spot structure: s_G; bonds 1–5.

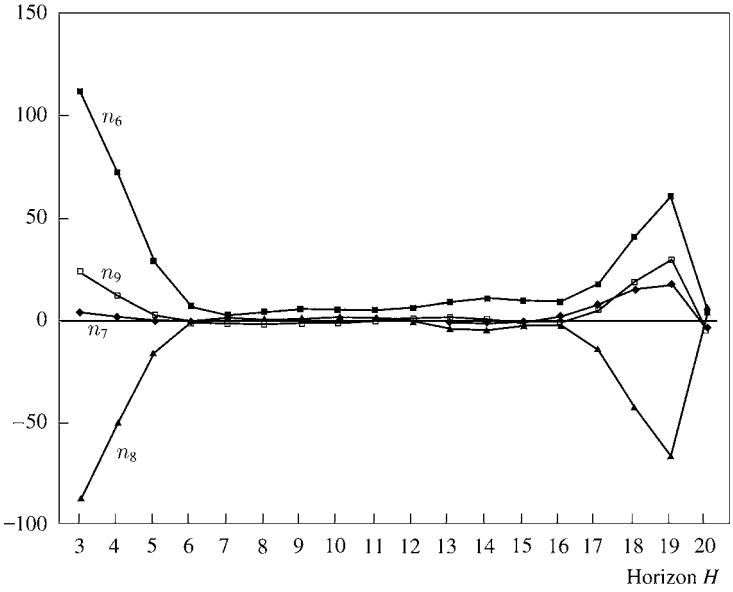


Fig. 18. Immunizing portfolios as a function of horizon H; initial spot structure: s_G; bonds 6–9.

costs could be achieved, for instance, by studying numerically the relationship between the immunization parameter K and the transaction cost. Indeed, starting from an initial, observed, spot rate structure, to each value of the arbitrary, positive, parameter K corresponds a given immunizing portfolio, which entails a given transaction cost. This could be done when building up a portfolio from scratch. But it could be done also in order to transform an existing, non-protected, portfolio into an immunized one at minimal cost.

6. Notes to references

The history of immunization can be found in the book edited by Hawawini (1982a). The reader will find in it not only the cornerstones of duration analysis (the seminal papers by Macaulay (1938), Hicks (1939) and Redington (1952)), but a detailed analysis of the Macaulay duration by Hawawini (1982b). The book also carries important articles by Hopewell and Kaufman (1973), Ingersoll, Skelton and Weil (1978), Fisher and Weil (1977), Bierwag (1977). It is to be noted that the duration concept was rediscovered independently by Hicks (1939) and Samuelson (1945).

A set of important, recent papers on the subject can be found in the book edited by Nawalkha and Chambers (1999a, 1999b). Particularly of note for our subject are the papers by Nawalkha (1999a, 1999b), Nawalkha and Chambers (1999a, 1999b), Chambers, Carleton and McEnally (1999). The application of the powerful model by Heath, Jarrow and Morton (1992) to bond analysis has been carried out in Jarrow and Turnbull (1996). A detailed account and applications to immunization is available in de La Grandville (2001). A deep analysis of the mathematics of immunization is developed in Pakes (2001) in the general framework of bonds that may pay continuous cash flows.

We should mention other, very powerful, methods of dealing with bond management, based upon recent advances in stochastic programming. The chapter in this volume by J. Dupacova, M. Bertocchi and V. Moriggio "Bond portfolio management via stochastic programming" gives the state of the art in this field. Also, the reader should be aware that immunization strategies are usually set up using linear programming models. The optimization models based upon linear programming are discussed in Zenios (1993). Finally, we should stress here that we have dealt with AAA, i.e., defaultless, bond immunization; so only interest rate risk is taken into consideration. If a portfolio manager wants to immunize a portfolio that includes corporate bonds, he should definitely consider factor immunization. Recent advances on the subject are in the paper by M. Bertocchi, R. Giacometti and S.A. Zenios (2005).

References

Bertocchi, M., Giacometti, R., Zenios, S.A., 2005. Risk factor analysis and portfolio immunization in the corporate bond market. European Journal of Operational Research 161, 348–363.

Bierwag, G., 1977. Immunization, duration and the term structure of interest rates. Journal of Financial and Quantitative Analysis, December.

Chambers, D., Carleton, W., McEnally, R., 1999. Immunizing default-free bond portfolios with a duration vector. Journal of Financial and Quantitative Analysis.

de Boor, C., 1978. A Practical Guide to Splines. Springer-Verlag, New York.

Fisher, L., Weil, R., 1977. Coping with the risk of interest rate fluctuations: returns to bondholders from naïve and optimal strategies. Journal of Business, 408–431, July.

Hawawini, G., 1982a. Bond Duration and Immunization: Early Developments and Recent Contributions. Garland, New York.

Hawawini, G., 1982b. On the mathematics of bond duration. In: Hawawini, G. (Ed.), Bond Duration and Immunization: Early Developments and Recent Contributions. Garland, New York, pp. 1–9.

Heath, D., Jarrow, R., Morton, A., 1992. Bond pricing and the term structure: a new methodology for contingent claim valuations. Econometrica 60 (1), 77–105, January.

Hicks, J., 1939. Value and Capital. Clarendon Press, Oxford, pp. 185–188.

Hopewell, M., Kaufman, G., 1973. Bond price volatility and the term to maturity: a generalized respecification. American Economic Review, 749–753, September.

Ingersoll, J., Skelton, M., Weil, R., 1978. Duration forty years later. Journal of Financial and Quantitative Analysts, November.

Jarrow, R., Turnbull, S., 1996. Derivative Securities. South Western College Publ., Cincinnati, OH.

de La Grandville, O., 2001. Bond Pricing and Portfolio Analysis—Protecting Investors in the Long Run. MIT Press, Cambridge, MA.

Macaulay, F., 1938. Some Theoretical Problems Suggestment by the Movement of Interest Rates, Bond Yield, and Stock Prices in the United States Since 1856. Columbia University Press, New York, pp. 44–53.

Nawalkha, S., 1999a. Duration sixty years later. In: Chambers, D., Nawalkha, S. (Eds.), Interest Rate Risk Measurement and Management. Institutional Investor Journals.

Nawalkha, S., 1999b. The M-vector: derivation and testing of the extensions to M-square. In: Chambers, D., Nawalkha, S. (Eds.), Interest Rate Risk Measurement and Management. Institutional Investor Journals.

Nawalkha, S., Chambers, D. (Eds.), 1999a. Interest Rate Risk Measurement and Management. Institutional Investor Journals.

Nawalkha, S., Chambers, D., 1999b. An improved immunization strategy: the M-absolute. In: Chambers, D., Nawalkha, S. (Eds.), Interest Rate Risk Measurement and Management. Institutional Investor Journals.

Pakes, A., 2001. Behaviour of bond portfolios immunized against polynomial variations of interest rates, Research Paper, Department of Mathematics and Statistics, University of Western Australia.

Redington, F., 1952. Review of the principles of life-office valuations. Journal of the Institute of Actuaries 8, 286–340.

Samuelson, P.A., 1945. The effect of interest rate increases on the banking system. The American Economic Review, 16–27, March.

Wahba, G., 1990. Spline Models for Observational Data. SIAM, Philadelphia.

Zenios, S.A. (Ed.), 1993. Financial Optimization. Cambridge University Press, Cambridge, England.

Chapter 4

RISK-RETURN ANALYSIS

HARRY M. MARKOWITZ

Harry Markowitz Company, 1010 Turquoise Street Suite 245, San Diego, CA 92109, USA

ERIK VAN DIJK

Compendeon b.v., Kosterijland 48, 3981 AJ Bunnik, P.O. Box 62, 3980 CB Bunnik, The Netherlands
e-mail: erik.vandijk@compendeon.com

Contents

Abstract 140
Keywords 141
1. Introduction 142
2. The "general" mean-variance model 144
3. Applications of the general model 146
   3.1. Asset liability modeling 146
   3.2. Factor models 147
   3.3. Other constraints 148
   3.4. Tracking error 148
4. Examples of mean-variance efficient sets 149
   4.1. Critical lines and corner portfolios 149
   4.2. Efficient EV and $E\sigma$ combinations 150
   4.3. All feasible $E\sigma$ combinations 152
   4.4. Possible features 153
5. Solution to the "general" mean-variance problem 156
   5.1. Preliminaries 156
   5.2. The critical line algorithm 156
   5.3. Getting started 158
   5.4. The critical line algorithm with upper bounds 161
   5.5. The critical line algorithm with factor and scenario models of covariance 162
   5.6. Notes on computing 165
6. Separation theorems 166
   6.1. The Tobin–Sharpe separation theorems 166
   6.2. Two-funds separation 169
   6.3. Separation theorems not true in general 169

Handbook of Asset and Liability Management, Volume 1
Edited by S.A. Zenios and W.T. Ziemba
Copyright © 2006 Elsevier B.V. All rights reserved
DOI: 10.1016/S1872-0978(06)01004-0

6.4. The Elton, Gruber, Padberg algorithm 170
6.5. An alternate EGP-like algorithm 171
7. Alternate risk measures 173
7.1. Semideviation 173
7.2. Mean absolute deviation (MAD) 175
7.3. Probability of loss and value at risk (Gaussian R_p) 176
7.4. Probability of loss and Value at Risk (non-Gaussian R_p) 178
7.5. Conditional value at risk (CVaR) 180
8. Choice of criteria 180
8.1. Exact conditions 180
8.2. Mean-variance approximations to expected utility 181
8.3. Significance of MV approximations to EU 184
9. Risk-return analysis in practice 184
9.1. Choice of criteria 185
9.2. Tracking error or total variability 186
9.3. Estimates for asset classes 187
9.4. Estimation of expected returns for individual equities 187
9.5. Black–Litterman 188
9.6. Security analyst recommendations 189
9.7. Estimates of covariance 189
9.8. Parameter uncertainty 190
10. Epilogue 192
References 193

Abstract

A risk-return analysis seeks "efficient portfolios", i.e., those which provide maximum return on average for a given level of portfolio risk. It examines investment opportunities in terms familiar to the financial practitioner: the risk and return of the investment portfolio. On the other hand, it can be related to the theory of rational behavior over time and under uncertainty—e.g., as formulated by Von Neumann and Morgenstern, Leonard J. Savage and R. Bellman—by viewing a function of portfolio "risk" and "return" as an approximation to a "derived utility function". This view is useful in exploring practical questions such as choice among alternate measures of risk in the risk-return analysis, and the relationship between the single-period analysis and the many-period investment problem.

This chapter examines several risk-return criteria, but focuses principally on the oldest and still most widely used, namely, mean-variance analysis. Important properties and fast computing procedures are known for mean-variance analysis in general. Additional properties and much faster algorithms are known for certain of its distinguished special cases. This article explores both the general case and some of the distinguished special cases.

Keywords

mean-variance, semivariance, EGP algorithm, critical line algorithm, separation theorems, VaR, CVaR, mean-probability of loss, expected utility, choice of criteria

JEL classification: G11

1. Introduction

Roy (1952) and Markowitz (1952) proposed mean and variance of return on the investor's portfolio-as-a-whole as criteria for investment decision making. Markowitz proposed presenting the investor with a curve, now called the "efficient frontier", which shows the maximum feasible expected return which can be obtained for various levels of "risk", or minimum "risk" which can be obtained for various levels of expected return, where "risk" is measured by the variance or, equivalently, standard deviation of portfolio return. The investor is to choose a risk-return combination from this risk-return tradeoff curve according to his or her willingness to bear more risk for greater expected return.

The Roy proposal recommended the specific (mean-variance efficient) portfolio which maximizes

$$\frac{E - d}{\sigma} \tag{1.1}$$

for portfolio expected return E, standard deviation σ and some disastrous level of return d. Roy refers to this as the "safety first" portfolio.

In maximizing (1.1), Roy chose a portfolio subject to the constraint

$$\sum_{i=1}^{n} X_i = 1 \tag{1.2a}$$

where X_i is the "fraction" invested in the ith of n securities. Markowitz (1952) requires these fractions to be nonnegative

$$X_i \geqslant 0, \quad i = 1, \ldots, n. \tag{1.2b}$$

Markowitz (1956) presents properties and formulas for efficient frontiers when the constraint set consists of any collection of linear equalities and/or inequalities in variables which may or may not have nonnegativity constraints. Markowitz (1959), Appendix A generalizes Markowitz (1956) by allowing an arbitrary, possibly singular, covariance matrix.

We shall refer to (1.2a) by itself as "Roy's" constraint set, and (1.2a) plus (1.2b) as the "standard" constraint set. Tobin (1958) analyzes the market for "monetary assets" assuming that all investors seek mean-variance efficiency subject to the standard constraint set. Specifically, he assumes that one of the securities, "cash"—a.k.a. the "risk-free asset"—has zero variance. The remaining securities are "risky" with a positive definite covariance matrix. He shows that any portfolio which contains any cash at all consists of a mixture of cash plus a specific combination of risky securities. We will refer to this combination as the "tangent portfolio". The low risk portion of the efficient frontier consists of mixtures of cash and the tangent portfolio. Efficient portfolios with higher risk and return than the tangent portfolio contain no cash.

The formal model of Sharpe (1964) and Lintner (1965) differs from that of Tobin in that they assume that investors can borrow all they want, as well as lend all they have,

at the risk-free rate. They conclude that all efficient portfolios consist of either (a) one specific tangent portfolio, (b) the tangent portfolio plus cash, or (c) the tangent portfolio leveraged at the risk-free rate.

Two essential distinctions must be made between the objectives and content of Markowitz (1952, 1959), Tobin (1958), Sharpe (1964) and Lintner (1965). One is the distinction between "positive" and "normative" analyses. The former presents hypotheses about how the world (e.g., investors or markets) do in fact behave; the second presents maxims which recommend how investors ought to behave. Markowitz (1952) presented mean-variance analysis "both as a hypothesis to explain, and as a maximum to guide investment behavior". Thus it was both positive and normative in intent. In contrast, Markowitz (1959) presented mean-variance as a normative guide, not necessarily as a description of investor behavior. Tobin's model, and the Capital Asset Pricing Models of Sharpe (1964), Lintner (1965), Mossin (1966), Black (1972), Merton (1973) and others are positive models which seek to explain how markets work.

The second distinction is between those positive models which seek to explain the behavior of a single investor, like the Markowitz (1952) hypothesis about investor choice, and those which seek to explain the aggregate behavior of a collection of investors, like the market for monetary assets by Tobin, or capital markets in general by the CAPMs.

The scope of the present chapter is essentially that of Markowitz (1959)—normative analysis for a single (individual or institutional) investor. The typical reader is visualized as someone who will participate in a risk-return analysis in practice, especially someone who will help specify the details of a particular analysis. CAPM will be discussed here only incidentally. It is assumed that the reader already knows the definitions of mean (or expected value), variance, standard deviation, covariance, and the formula for the mean and variance of a weighted sum of random variables.[1] It is also assumed that the reader is familiar with matrix notation.[2]

Mean-variance analysis is an example of risk-return analysis with variance or, equivalently, standard deviation as the measure of risk. Other measures of risk which have been used in risk-return analysis include:

- semideviation (a.k.a. "downside" deviation);
- mean absolute deviation;
- value at risk (VaR);
- conditional value at risk (CVaR); and
- probability of falling below some level.

Any of these could be the risk measure in a risk-return analysis. We shall primarily discuss mean-variance, but note the extent to which our analyses extend to risk-return analyses using other measures of risk. Some of these are discussed in other chapters of this volume.

The remainder of this chapter is organized as follows: Section 2 defines the "general" mean-variance model; Section 3 illustrates applications of this model, e.g., to

[1] See Feller (1970), Chapter IX, or Markowitz (1959), Chapters 3 and 4.

[2] See Markowitz (1959), Chapter 8, or an introductory text on matrix algebra such as Aitken (1956).

asset/liability management; Section 4 illustrates the solution of the model; Section 5 discusses the solution in general; Section 6 discusses "separation theorems" including conditions under which they do or do not obtain; Section 7 discusses efficient frontiers for alternate measures of risk; Section 8 discusses choice of criteria; Section 9 surveys methods for estimating the parameters needed for a risk-return analysis; Section 10 is an Epilogue.

2. The "general" mean-variance model

Suppose that the return R_p on the portfolio-as-a-whole over some forthcoming period is a weighted sum of the n security returns $R' = (r_1, \ldots, r_n)$,

$$R_p = R'X \tag{2.1}$$

where the weights $X' = (X_1, \ldots, X_n)$ are chosen by the investor. We assume here that the r_i are jointly distributed random variables.

Therefore, also assuming that r_i have finite means and variances,

$$E_P = \sum_{i=1}^{n} \mu_i X_i = \mu' X, \tag{2.2a}$$

$$V_p = \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij} X_i X_j = X' C X \tag{2.2b}$$

where E_p, V_p are the expected return and variance of the portfolio-as-a-whole, $\mu' = (\mu_1, \ldots, \mu_n)$ the expected return on the n securities, σ_{ij} the covariance between r_i and r_j, and C the covariance matrix (σ_{ij}). In particular,

$$\sigma_{ii} = \sigma_i^2 = V_i$$

where σ_i is the standard deviation and V_i the variance of r_i. $\sigma_p = \sqrt{V_p}$ is the standard deviation of portfolio return.

The "general" (single period mean-variance) portfolio selection problem of Markowitz (1956) seeks the set of all efficient $E_p V_p$ or $E_p \sigma_p$ combinations, and efficient portfolios which have these mean-variance combinations, where portfolios are chosen subject to any system of linear equality

$$\sum_j a_{ij} X_j = b_i, \quad i \leqslant k_1, \tag{2.3a}$$

and/or linear inequality constraints

$$\sum_j a_{ij} X_j \geqslant b_i, \quad k_1 < i \leqslant k_1 + k_2, \tag{2.3b}$$

$$\sum_j a_{ij} X_j \leqslant b_i, \quad k_1 + k_2 < i \leqslant k_1 + k_2 + k_3, \tag{2.3c}$$

and where none, some or all the X_i are required to be nonnegative

$$X_i \geqslant 0, \quad i \leqslant n_1 \leqslant n. \tag{2.3d}$$

E.g., in Markowitz (1952), $k_1 = 1, k_2 = k_3 = 0, n_1 = n$; in Roy $k_1 = 1, k_2 = k_3 = 0$, $n_1 = 0$.

It can be shown[3] that for any model (M_1) with constraints as in (2.3) there is an equivalent model (M_2) whose constraints are written as

$$AX = b, \tag{2.4a}$$

$$X \geqslant 0 \tag{2.4b}$$

where A is a "constraint matrix" and b a "right-hand side" vector. Model M_2 is equivalent to M_1 in that it has the same set of efficient $E_p V_p$ combinations, and the corresponding efficient portfolios in M_2 can be easily mapped to efficient portfolios in M_1. To convert a model written in form (2.3) into an equivalent model written in form (2.4), change inequalities to equalities by introducing slack variables; e.g., write

$$\sum_{j=1}^{n} a_{ij} X_j + X_{is} = b_i, \quad X_{is} \geqslant 0,$$

for

$$\sum a_{ij} X_j \leqslant b_i.$$

Also, substitute

$$X_{iP} - X_{iN}$$

for any X_i not required to be nonnegative, where $X_{iP} \geqslant 0$, $X_{iN} \geqslant 0$ are respectively the "positive part" and "negative part" of X_i.

In describing applications it is often convenient to refer to equalities and/or inequalities as in (2.3). In describing computational procedures it is convenient to use the equivalent formulation in (2.4).

The general (single period) mean-variance portfolio selection problem, then, is equivalent to finding mean-variance efficient portfolios given arbitrary μ and C in (2.2) and arbitrary A and b in (2.4). In particular, it is not required that C be nonsingular. This is important since X may include risk-free securities, slack variables and pairs of securities which represent short and long positions. Also, sometimes C is estimated from historical returns with less periods (e.g., months of observation) than there are securities. Any of these circumstances will result in $|C| = 0$.

[3] See Markowitz and Todd (2000) (henceforth, MT), Chapter 2.

3. Applications of the general model

3.1. Asset liability modeling

The variables $X_1, \ldots, X_n$ of the general model (2.3) or (2.4) may represent assets, liabilities or other related variables, as illustrated in this section. To start, consider the inclusion of liabilities as well as assets in the model. For example, let us modify the standard model (1.2a), (1.2b) by permitting borrowing at a rate r_b. The constraint set may be written

$$\sum_{i=1}^{n} X_i - X_{n+1} = 1, \tag{3.1a}$$

$$X_i \geqslant 0, \quad i = 1, \ldots, n+1. \tag{3.1b}$$

If borrowing is limited to an amount no greater than u_b, then

$$X_{n+1} \leqslant u_b.$$

In this model, $\mu_{n+1} = -r_b$, $\sigma_{n+1} = 0$.

If the investor can borrow up to u_1 at the rate r_b and an additional u_2 at a higher rate r_c, this can be modeled as follows

$$\sum_{i=1}^{n} X_i - X_{n+1} - X_{n+2} = 1, \tag{3.2a}$$

$$X_i \geqslant 0, \quad i = 1, \ldots, n+2, \tag{3.2b}$$

$$X_{n+k} \leqslant u_k, \quad k = 1, 2, \tag{3.2c}$$

$$\mu_{n+1} = -r_b, \quad \mu_{n+2} = r_c, \quad \sigma_{n+1} = \sigma_{n+2} = 0.$$

The optimizer will figure out that it should not use any of X_{n+2} unless $X_{n+1} = u_1$. The more specific models (3.1) and (3.2) are special cases of the general model (2.3) or, equivalently, (2.4).

Next consider a pension plan whose investments are subject to the constraint

$$\sum_{i=1}^{n} X_i = d, \tag{3.3a}$$

$$X_i \geqslant 0, \quad i = 1, \ldots, n, \tag{3.3b}$$

where d is the dollar value of assets available for investment. Here it is convenient to let the decision variable X_i represent dollar investment rather than percent invested. (The general model imposes no requirement that the X_i sum to 1.0.) So far (3.3) is equivalent to the standard model. As before let $r_1, \ldots, r_n$ represent the return in the forthcoming "year" on the n available investments. In addition, suppose that $-r_{n+1}$ represents the increase or decrease during the year in the present value of the plan's future benefit obligations. This may depend on random forces such as interest rate changes

or salary inflation. The investment decision will not consider altering benefits but, for given expected return on the investment portfolio, it will seek to minimize the variance of investment returns *less* obligation changes:

$$\sum_{i=1}^{n} X_i r_i + r_{n+1} \tag{3.3c}$$

where r_{n+1} is *minus* the increase in obligation changes. This is accomplished by introducing a "fictitious" investment X_{n+1} and requiring

$$X_{n+1} = 1. \tag{3.3d}$$

X_{n+1} is not introduced into (3.3a). The $(n+1)$ by $(n+1)$ covariance matrix C now includes the covariances

$$\sigma_{i,n+1}, \quad i = 1, \ldots, n,$$

between the various investments and minus the change in obligations. $\sigma_{n+1,n+1}$ is included to keep C positive semidefinite, but we may let $\mu_{n+1} = 0$ and interpret E as the return of the investment assets A. See Sharpe and Tint (1990) for further details including the recommendation that (3.3c) be replaced by

$$\sum_{i=1}^{n} X_i r_i + k r_{n+1} \tag{3.3e}$$

for $0 \leqslant k \leqslant 1.0$. With $k = 0$ (3.3e) provides "asset only" analysis; with $k = 1.0$ (3.3e) provides what Sharpe and Tint call "full surplus optimization". They suggest circumstances under which $0 < k < 1.0$ is desirable. This can be incorporated into the general model by reinterpreting "r_{n+1}" in (3.3c).

3.2. Factor models

The modeling of covariance gives rise to another kind of variable and constraint. The one-factor model, Sharpe (1963), will illustrate. Suppose that the return on the ith security is

$$r_i = \alpha_i + \beta_i F + u_i \tag{3.4}$$

where α_i and β_i are constants, F is a random common factor and the u_i are uncorrelated with F and each other. Given estimates of β_i, $\mathrm{Var}(F)$ and $V(u_i)$ one can proceed in either of two directions. It can be shown that

$$\mathrm{cov}(r_i, r_j) = \beta_i \beta_j \,\mathrm{Var}(F) \quad \text{for } i \neq j, \tag{3.5a}$$

$$V(r_i) = \beta_i^2 \,\mathrm{Var}(F) + V(u_i) \quad \text{for } i = 1, \ldots, n. \tag{3.5b}$$

The covariance matrix C can be computed from (3.5) and this can be used in the computation of efficient portfolios.

Alternatively, we note that the return on the portfolio-as-a-whole is

$$R_p = \sum X_i r_i = \left(\sum \alpha_i X_i\right) + \left(\sum \beta_i X_i\right) F + \sum X_i u_i$$
$$= A + BF + \sum X_i u_i. \tag{3.6}$$

Hence

$$V(R_p) = B^2 V(F) + \sum X_i^2 V(u_i) \tag{3.7}$$

where

$$B = \sum_{i=1}^{n} \beta_i X_i. \tag{3.8}$$

Thus if we add an $(n + 1)$st security to the portfolio analysis, X_{n+1} ($\equiv B$ above), plus one linear constraint (3.8), the variance of the portfolio (3.7) becomes a sum of squares in $n + 1$ variables. This has great computational advantages as compared to tracing out mean-variance frontiers with a C matrix dense with nonzero entries.

Similar advantage (albeit in some cases not as great) can be had by introducing suitable fictitious variables and constraints into multifactor models (Cohen and Pogue, 1967), scenario or mixed scenario and factor models (Markowitz and Perold, 1981a, 1981b) and historical models when there are many more securities than time periods (Markowitz et al., 1993a, 1993b).

3.3. Other constraints

The constraints permitted by (2.3) or (2.4) can be used to limit exposure to various securities, industries or sectors, or limit how much the new portfolio differs from the previous one, referred to as a "turnover constraint".

If returns were certain, the portfolio problem would be to maximize the linear function (2.1) subject to the constraints (2.3) or (2.4). That is, it would be a "linear programming" problem. (See Dantzig (1963), Koopmans (1951).) Modeling systems using linear programming constraint sets has been widely explored. The general portfolio selection problem may be viewed as linear programming with random payoffs and with portfolio mean and variance as criteria.

3.4. Tracking error

In practice, money managers frequently seek portfolio return R_p which outperforms return R_B on some benchmark such as the Standard and Poor's (S&P) 500 index of stock prices. In a mean-variance framework they seek high expected excess return

$$E(R_p - R_B) = E(R_p) - E(R_B) \tag{3.9a}$$

for given variance of excess return:

$$V(R_p - R_B) = V(R_p) + V(R_B) - 2\,\mathrm{cov}(R_p, R_B). \tag{3.9b}$$

In particular, the covariance of securities with the benchmark affect the last term of (3.9b) and therefore the composition of efficient portfolios. Typically the benchmark is the return on a portfolio of securities X_B which changes from time to time. As a rule it is better to estimate covariances with the current benchmark portfolio X_B rather than use historical estimates of covariances with R_B reflecting changing X_B over time. This can be accomplished, if X_B is known, by defining the difference

$$X_D = X_p - X_B \tag{3.9c}$$

where X_p is the actual (contemplated) portfolio. Then (3.9a) and (3.9b) become

$$E(R_p - R_B) = X_D'\mu, \tag{3.9a$'$}$$

$$V(R_p - R_B) = X_D' C X_D, \tag{3.9b$'$}$$

E and V are now expressed in terms X_D only. X_p can be retained along with X_D as variables, and (3.9c) added as constraints. Or, (3.9c) can be used to replace X_p by X_D in (2.4). Either way the new portfolio problem remains a special case of the general portfolio selection problem.

4. Examples of mean-variance efficient sets

4.1. Critical lines and corner portfolios

Three-security examples illustrate many of the properties of general mean-variance efficient sets. To begin, let

$$\mu = (0.16, 0.13, 0.12),$$
$$\sigma = (\sigma_1, \sigma_2, \sigma_3) = (0.2, 0.25, 0.2).$$

For convenience of hand computation, if you care to check us, in this first example we assume $\sigma_{ij} = 0$ for $i \neq j$.

First consider the Roy constraint set with (1.2a) as the only constraint. To minimize V (or, equivalently, $V/2$) for given E subject to (1.2a), define the Lagrangian

$$L = V/2 - \lambda_E\Big(\sum X_i \mu_i\Big) + \lambda_A\Big(\sum X_i\Big). \tag{4.1}$$

Set $\partial L/\partial X_i$ to zero for each i, append (1.2a) and rewrite as

$$\begin{pmatrix} 0.04 & 0 & 0 & 1 \\ 0 & 0.0625 & 0 & 1 \\ 0 & 0 & 0.04 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \lambda_A \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} + \lambda_E \begin{pmatrix} 0.16 \\ 0.13 \\ 0.12 \\ 0 \end{pmatrix}. \tag{4.2a}$$

As written, (4.2a) minimizes $V/2 - \lambda_E \sum \mu_i X_i$ for fixed λ_E rather than minimizing $V/2$ for fixed E. This saves one equation here, and has other advantages. (See MT (2000), Chapter 6.) Solve (4.2a) to obtain

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \lambda_A \end{pmatrix} = \begin{pmatrix} 0.3788 \\ 0.2424 \\ 0.3788 \\ -0.0152 \end{pmatrix} + \begin{pmatrix} 0.5606 \\ -0.1212 \\ -0.4394 \\ 0.1376 \end{pmatrix} \lambda_E. \tag{4.2b}$$

If we drop the last row in (4.2b) we have a straight line in portfolio space with the Lagrangian multiplier λ_E as parameter. The projection of this line onto the X_1X_2-space is the line LL' passing through points e, c, d in Figure 1(a). We shall refer to this as the "critical line" of the three-security model. Since V is convex, the points on the line (4.2b) minimize L for given λ_A, λ_E. This implies that they minimize V among portfolios with the same E and $\sum X_i$.

For the standard constraint set (1.2a), (1.2b), the points on and in the triangle Oab in Figure 1(a) are the projection on the X_1X_2-space of the set of feasible portfolios. In the example, the undiversified portfolio at the point a, with $X_1 = 1$, has maximum feasible expected return. We will refer to the expected value of the portfolios at $a, b, c, \ldots$ as $E_a, E_b, E_c, \ldots$. Points on the critical line LL' with E between E_c and E_d are feasible and therefore efficient, as they are with the Roy constraint set. For an $E > E_d$ the point of LL' with this E is not feasible for the standard model. For any $E \in [E_d, E_a]$ the point with this E on the segment $\overline{da}$ is efficient.

Thus the set of efficient portfolios for the standard constraint set in this example is the piecewise linear set in portfolio space whose projection on X_1X_2-space is the piecewise linear set cda. The points c, d and a are referred to as *corner portfolios*.

4.2. *Efficient EV and Eσ combinations*

For any number of securities, a straight line in portfolio space can be written as

$$X = a + b\theta \tag{4.3}$$

where a and b are n-component column vectors, $b \neq 0$, and θ a scalar parameter. Substituting (4.3) into (2.2) we get

$$E_p = a'\mu + (b'\mu)\theta = d_0 + d_1\theta, \tag{4.4a}$$

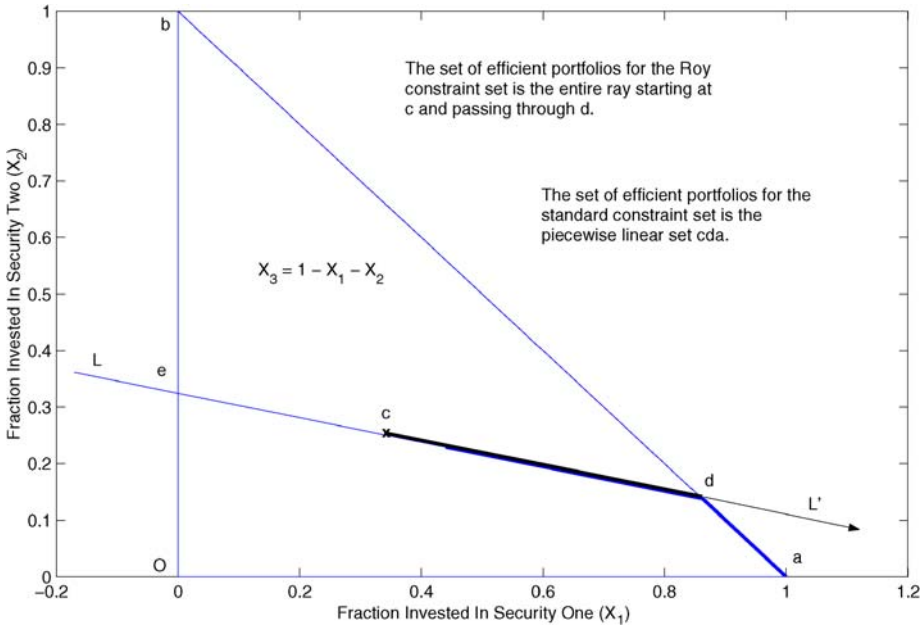
$$V_p = a'Ca + 2(a'Cb)\theta + (b'Cb)\theta^2 = \tilde{c}_0 + \tilde{c}_1\theta + \tilde{c}_2\theta^2. \tag{4.4b}$$

If $b'\mu \neq 0$, i.e., as long as b is not a "zero-mean direction", we can express θ in terms of E_p in (4.4a), and substitute this into (4.4b) to obtain

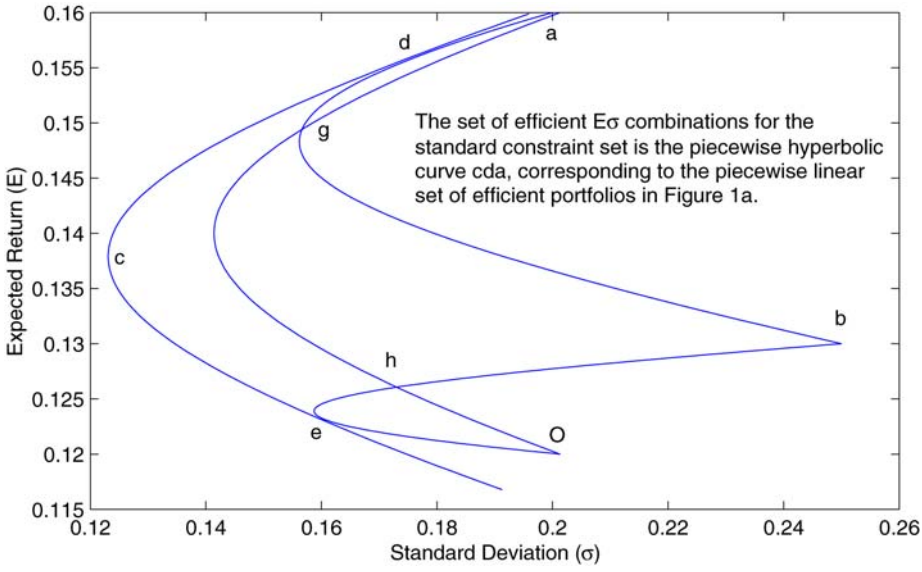
$$V_p = c_0 + c_1E_p + c_2E_p^2. \tag{4.4c}$$

Completing the square we can rewrite this as

$$V_p = V_{\text{Min}} + c_2(E_p - E_{\text{Min}})^2. \tag{4.4d}$$



(a)



(b)

Fig. 1. (a) Three security example portfolio space. (b) Three security example E, σ space.

Since $V_p \geqslant 0$ always, we have $V_{\text{Min}} \geqslant 0$. In general, if b is not a "zero-variance direction", i.e., if $0\ b'Cb > 0$, e.g., if C is positive definite as in the example, then V_p as a function of E_p is a parabola which reaches a minimum V_{Min} at $E_p = E_{\text{Min}}$. Portfolio standard deviation σ_p satisfies

$$\sigma_p^2 = V_{\text{Min}} + c_2(E_p - E_{\text{Min}})^2. \tag{4.5a}$$

If $V_{\text{Min}} = 0$ then (4.5a) is satisfied by

$$\sigma_p = \pm\sqrt{c_2}|E_p - E_{\text{Min}}|. \tag{4.5b}$$

Since only $\sigma_p \geqslant 0$ has meaning for us, the solution when $V_{\text{Min}} = 0$ is the pair of rays which satisfy

$$\sigma_p = \sqrt{c_2}|E_p - E_{\text{Min}}|. \tag{4.5c}$$

When $V_{\text{Min}} > 0$ then σ_p as a function of E_p in (4.5a) is a hyperbola with the pair of lines in (4.5b) as asymptotes. Only the branch of the hyperbola with $\sigma_p > 0$ has meaning for us.

Markowitz (1952) plotted risk (e.g., V_p) as a function of E_p with E_p on the horizontal axis and risk on the vertical. Roy plotted E_p on the vertical and risk on the horizontal. The latter is now the common practice, and we will therefore conform to it.

The curve passing through points d, c, e in Figure 1(b), extended without bounds in both directions, is the branch of the hyperbola which relates σ_p to E_p on the critical line LL'. The set of efficient $E_p\sigma_p$ combinations for the Roy constraint set in this example is the "half" of the hyperbola that starts at $E_{\text{Min}}\sigma_{\text{Min}}$ ($\sigma_{\text{Min}} = \sqrt{V_{\text{Min}}}$ of course) and moves in the direction of increasing E_p.

In $E\sigma$ space, the hyperbola *adgb* in Figure 1(b) relates σ to E on the line *ab* which has $X_3 = 0$. This hyperbola touches the *adce* hyperbola at the point d, in Figure 1(b), corresponding to the corner portfolio d in Figure 1(a) where the two pieces of the efficient set meet. In $E\sigma$ space, the set of efficient $E\sigma$ combinations is the piecewise hyperbolic curve *cda*.

4.3. All feasible $E\sigma$ combinations

One may have occasion to draw the set of all feasible $E\sigma$ combinations, perhaps to contrast them with efficient combinations. One should therefore know what this set looks like. The portfolios which *maximize* V for given E lie on the boundary of the feasible set; in particular, in standard three security examples the portfolio which maximizes V for a given feasible E_0 is on a boundary line of the feasible set $\overline{Oa}$, $\overline{Ob}$, or $\overline{ab}$. The hyperbolas labeled *Oa*, *Ob* and *ab* respectively in Figure 1(b) are the E, σ curves for the portfolios on the corresponding lines. Maximum feasible σ as a function of E, in the range from the highest to the lowest feasible E, follows the *agO* hyperbola until point g, then follows the *ab* hyperbola until the point b, then the *bO* hyperbola until point h where it again follows the *aghO* hyperbola. This entails jumps from one boundary line

to another in the portfolio space of Figure 1(a). No such discontinuity exists in the set of portfolios which *minimize* σ for various E.[4]

4.4. Possible features

Figures 2(a) through (c) illustrate some ways in which a three security example can differ from that in Figure 1. The means, variances and covariances for these examples are presented in Table 1. The examples were made up to illustrate certain mathematical possibilities. But the parameters in Table 1 are not outlandish and any of these possibilities could conceivably arise in practice. In the example in Figure 1, the portfolio c which minimizes σ_p subject only to the budget constraint (1.2a) is feasible, i.e., satisfies (1.2b). This is inevitable in the standard model when security returns have positive σ_i and are uncorrelated. Specifically, the σ_p minimizing portfolio $\widehat{X}$ has $\widehat{X}_i = k/\sigma_i^2$ where $k = 1/(\sum 1/\sigma_i^2)$. When returns are correlated the σ_p minimizing portfolio c may or may not be feasible.

In particular, in case (a) of Figure 2 and Table 1 the portfolio c which minimizes σ_p subject only to (1.2a) lies outside the standard feasible set, while the feasible portfolio $X_{\min}$ which minimizes σ_p lies on the bO boundary of the feasible set. In this example the critical line enters the feasible set at e and leaves it at d, moving in the direction of increasing E_p. The set of efficient portfolios for the standard constraint set is the piecewise linear curve $X_{\min}eda$. The points $X_{\min}$, e, d and a are the corner portfolios of the efficient set.

In example (b) in Figure 2 and Table 1 the critical line does not intersect the feasible set. The efficient set starts at $X_{\min}$, moves from $X_{\min}$ to O, "turns the corner" at O, next moves along $\overline{Oa}$ ending at portfolio a with maximum E_p. In this case the set

Table 1
Parameters for Figure 2

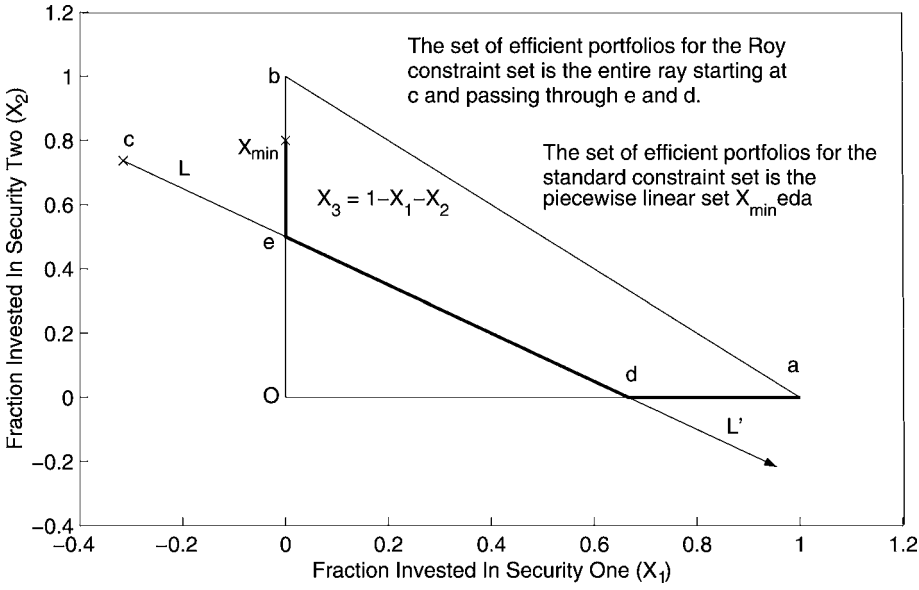
In each case
$\mu' = (0.16, 0.10, 0.13)$

$$C_a = \begin{pmatrix} 0.0650 & 0.0100 & 0.0350 \\ 0.0100 & 0.0100 & 0.0050 \\ 0.0350 & 0.0050 & 0.0250 \end{pmatrix}$$

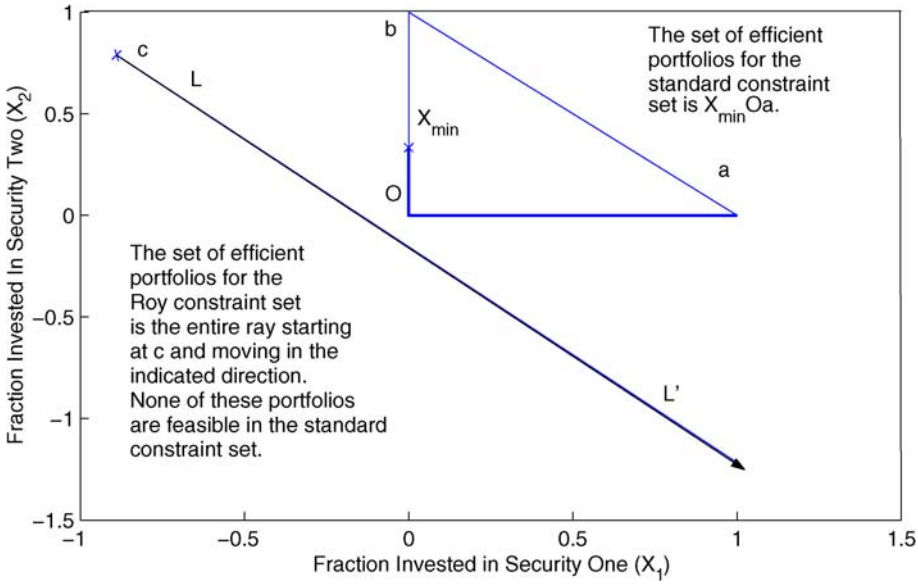
$$C_b = \begin{pmatrix} 0.0500 & 0.0250 & 0.0250 \\ 0.0250 & 0.0200 & 0.0100 \\ 0.0250 & 0.0100 & 0.0150 \end{pmatrix}$$

$$C_c = \begin{pmatrix} 0.1035 & 0.0191 & 0.0541 \\ 0.0191 & 0.0108 & 0.0163 \\ 0.0541 & 0.0163 & 0.0343 \end{pmatrix}$$

[4] When the set of portfolios which minimize V for given E is unique. When the set is not unique, there always exists a solution, i.e., a set of efficient portfolios—one for each efficient $E\sigma$ combination—without discontinuities.

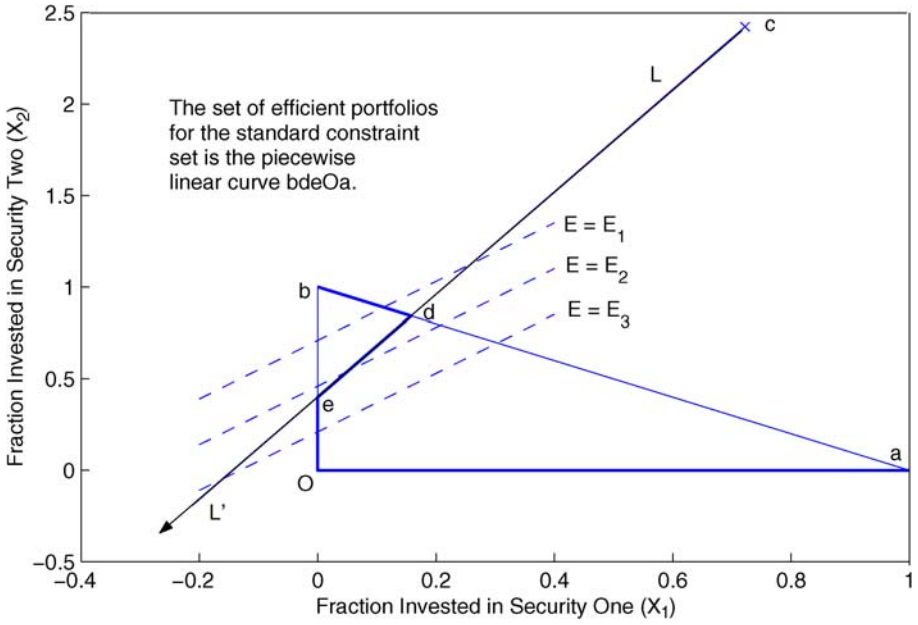


(a)



(b)

Fig. 2. (a) Three security example with "unconstrained" min V not feasible. (b) Three security example in which the critical line LL' does not intersect the feasible set. (c) Three security example whose efficient set has the maximum number of segments.



(c)

Fig. 2. (*Continued.*)

of efficient EV combinations and efficient $E\sigma$ combinations each have a "kink", i.e., a discontinuous first derivative, at the point O.

As discussed in the next section, when the set of efficient portfolios is unique, it is piecewise linear. On any of these linear pieces certain securities are "IN", the others are "OUT". The OUT securities have zero investment. Starting with the piece containing the V minimizing portfolio, in Figure 1 the IN sets are $\{1, 2, 3\}$, $\{1, 2\}$ and $\{1\}$; in Figure 2(a) they are $\{2, 3\}$, $\{1, 2, 3\}$, $\{1, 3\}$ and $\{1\}$; in Figure 1(b) they are $\{2, 3\}$, $\{3\}$, $\{1, 3\}$ and $\{1\}$.

Figure 2(c) provides a case in which every nonempty subset of $\{1, 2, 3\}$ is the IN set on some efficient segment, namely, starting with the minimum V portfolio, the IN sets are $\{2\}$, $\{1, 2\}$, $\{1, 2, 3\}$, $\{2, 3\}$, $\{3\}$, $\{1, 3\}$, $\{1\}$. Since portfolio (b)—with $X_2 = 1.0$ therefore $E_p = 0.10$—has minimum variance, and portfolio (a)—with $X_1 = 1.0$ therefore $E_p = 0.16$—has maximum mean, there is an efficient portfolio for every $E \in [0.10, 0.16]$. The line labeled $E = E_1$, is an *isomean* line, the set of all portfolios satisfying (1.2a) but not necessarily (1.2b) with expected return equal some level E_1. The lines parallel to $E = E_1$, labeled $E = E_2$ and $E = E_3$, are isomean lines for other (higher) values of portfolio E. For $E_k \in [0.10, 0.16]$ the *efficient* portfolio on the isomean line $E = E_k$ is the *feasible* portfolio on the line closest to the critical line LL'.

We see from Figure 2(c) that for $E = E_1$ the efficient portfolio is on the segment $\overline{bd}$, for $E = E_2$ it is on the segment $\overline{de}$, and for $E = E_3$ it is on the segment $\overline{eO}$.

Since a given IN set can appear at most once in the piecewise linear set of efficient portfolios (when the efficient set is unique, e.g., when $|C| \neq 0$) the example in Figure 2(c) has the maximum possible number of "efficient segments". Since portfolio $(0, 1, 0)$ has minimum feasible variance, if we changed the example so that $\mu_2 > \mu_i$, $i = 1, 3$, then $X = (0, 1, 0)$ would be the only efficient portfolio and $\{2\}$ the only IN set.

5. Solution to the "general" mean-variance problem

5.1. Preliminaries

In many ways the solution to the general mean-variance problem is illustrated by the three-security examples just discussed. In this section we describe, mostly without proof, the principal features and formulas of the solution to the general problem. See Markowitz (1959) Appendix A, Perold (1984), Markowitz (1987) or MT (2000) for proofs and further details.

We begin by noting some "unusual" possibilities. We are given A, b, μ and C in (2.2) and (2.4). It is possible that, for the given A and b, no portfolio X satisfies (2.4). We say then that the model is *infeasible*. Otherwise it is *feasible*.

Feasibility is a property of the constraint set (2.4). Another property is whether the set of feasible portfolios is bounded (like the standard set defined by (1.2a), (1.2b)) or unbounded like the Roy set defined by (1.2a) only. If the constraint set is unbounded and C is singular, it is possible for a model to be feasible and yet have no mean-variance efficient portfolios. In this case, if $\widehat{X}$ is feasible with minimum V_p and with expected return E, there is another feasible portfolio X^* with the same V and with $E^* > E$. Below we assume that the model is feasible and has efficient portfolios.

5.2. The critical line algorithm

The Lagrangian expression for the general model is

$$L = V/2 + \sum_{k=1}^{m} \lambda_k \Big(\sum a_{kj} X_j \Big) - \lambda_E \sum \mu_i X_i. \tag{5.1}$$

Let

$$\eta = \left(\frac{\partial L}{\partial X_i} \right) = (CA'\mu) \begin{pmatrix} X \\ \lambda \\ -\lambda_E \end{pmatrix}. \tag{5.2}$$

For the moment, to develop a definition, arbitrarily select a nonempty subset of $\{1, 2, \ldots, n\}$ and designate this subset as the IN variables and its complement as the OUT variables. Let

$$M = \begin{pmatrix} C & A' \\ A & O \end{pmatrix} \tag{5.3}$$

and let M_{IN} be the M matrix with the rows and columns deleted that correspond to OUT variables. Similarly, let μ_{IN} and X_{IN} be the μ and X vectors with OUT components deleted, and 0_{IN} be a zero vector of the same size as μ_{IN}.

If M_{IN} is nonsingular we say that the arbitrarily chosen IN set has an associated *critical line* satisfying:

$$X_i = 0 \quad \text{for } i \in \text{OUT}$$

and

$$M_{\text{IN}} \begin{pmatrix} X_{\text{IN}} \\ \lambda \end{pmatrix} = \begin{pmatrix} 0_{\text{IN}} \\ b \end{pmatrix} + \begin{pmatrix} \mu_{\text{IN}} \\ 0 \end{pmatrix} \lambda_E \tag{5.4}$$

where λ are the m multipliers associated with the Lagrangian expression (5.1).

Multiplying through by M_{IN}^{-1} we can solve (5.4) for X_{IN} and λ as linear functions of λ_E

$$\begin{pmatrix} X_{\text{IN}} \\ \lambda \end{pmatrix} = \alpha_{\text{IN}} + \beta_{\text{IN}} \lambda_E. \tag{5.5a}$$

If we substitute (5.5a) into (5.2) we find that the η vector is also a linear function of λ_E:

$$\eta = \gamma_{\text{IN}} + \delta_{\text{IN}} \lambda_E. \tag{5.5b}$$

The critical line is constructed so that

$$X_i = 0 \quad \text{for } i \in \text{OUT}, \tag{5.6a}$$

$$\eta_i = 0 \quad \text{for } i \in \text{IN}. \tag{5.6b}$$

According to the above definition, in the 3-security case in particular, not only is the line LL' in Figure 1(a) the critical line (or, more precisely, the projection onto the X_1, X_2 plane of the critical line) for IN $= \{1, 2, 3\}$, but so are the X_1-axis, X_2-axis, the line $\{X_1, X_2\colon\ X_1 + X_2 = 1\}$, and the points $(1, 0)$, $(0, 1)$ and $(0, 0)$ with IN $= \{1, 3\}, \{2, 3\}, \{1, 2\}, \{1\}, \{2\}$ and $\{3\}$ respectively.

In light of (5.6a), (5.6b), if a point on the critical line also satisfies

$$X_i \geqslant 0 \quad \text{for } i \in \text{IN}, \tag{5.6c}$$

$$\eta_i \geqslant 0 \quad \text{for } i \in \text{OUT}, \tag{5.6d}$$

$$\lambda_E > 0 \tag{5.6e}$$

then the point is efficient, by the Kuhn and Tucker (1951) conditions. From (5.5) and (5.6), if any point on the critical is efficient then there will be an interval of that line (possibly open ended) all of whose points are efficient. We refer to such an interval as an *efficient segment*.

Since there are $2^n - 1$ nonnull subsets of $\{1, \ldots, n\}$ it is impractical to enumerate them all, determine which have nonsingular M_{IN}, among these determine which contain

efficient segments, then piece these together to form a complete set of efficient portfolios ("complete" in the sense of containing at least one feasible portfolio for every efficient $E\sigma$ combination). The *critical line algorithm* produces a complete, "nonredundant" set of efficient portfolios (i.e., a set with one and only one feasible portfolio for every efficient $E\sigma$ combination) and accomplishes this without searching among irrelevant IN sets. Note that if C is singular there may be more than one feasible (therefore efficient) portfolio that provides a given efficient $E\sigma$ combination. The complete nonredundant set of efficient portfolios produced by the critical line algorithm contains only one of these.

The critical line algorithm proceeds as follows: it traces out the efficient set from high to low E (and λ_E). For the moment let us postpone the question of how to get a first critical line with an efficient segment. We assume that we have a first or subsequent critical line with IN-set IN_t; we also assume we have in hand the corresponding M_{IN} and M_{IN}^{-1}. With the latter we can solve for α_{IN}, β_{IN}, γ_{IN}, δ_{IN} and with these in hand it is easy to determine which occurs first as λ_E is reduced:

$$X_i \downarrow 0 \quad \text{for } i \text{ IN}, \tag{5.7a}$$

$$\eta_i \downarrow 0 \quad \text{for } i \text{ OUT}, \tag{5.7b}$$

$$\text{or} \quad \lambda_E \downarrow 0. \tag{5.7c}$$

In case $\lambda_E \downarrow 0$ first, we have reached the efficient portfolio with minimum feasible V, and the algorithm stops. If C is singular, there may be more than one portfolio with minimum feasible V. Since the V minimizing portfolios may have different values of E, they may not all be efficient. But it is shown (e.g., in MT (2000)) that the portfolio reached by the critical line algorithm when $\lambda_E \downarrow 0$ is efficient as well as V minimizing.

If $X_i \downarrow 0$ first in (5.7), then i moves from IN to OUT on the next ("adjacent") efficient segment. On the other hand, if $\eta_i \downarrow 0$ first then i moves from OUT to IN in the new IN set, IN_{t+1}. (See MT (2000), Chapter 9 for what to do in case of ties for $X_i \downarrow 0$ and/or $\eta_i \downarrow 0$ prior to $\lambda_E \downarrow 0$.) If the algorithm has not stopped, because $\lambda_E \downarrow 0$ has not been reached, the new M matrix, $M_{\text{IN}(t+1)}$, is obtained from the old by adding or deleting one column and the corresponding row from the old $M_{\text{IN}(t)}$. This allows us to update M_{IN}^{-1} relatively inexpensively, and the process of using this to solve for α, β, γ, δ, etc. proceeds as before. This process repeats until $\lambda_E \downarrow 0$ is reached. It is shown that all IN sets reached by this process have nonsingular M_{IN}^{-1} and the end (with $\lambda_E \downarrow 0$) is reached in a finite number of iterations. Thus once a first efficient segment is found, each adjacent one can be determined in turn. (Actually, either the adjacent one with higher or lower λ_E can be determined from a given efficient segment, though we will proceed from high to low λ_E.)

5.3. Getting started

But how is a first IN set with nonsingular M to be obtained? We describe here how the procedure works when there is a unique portfolio which maximizes E. The general case is treated in MT (2000), Chapter 9.

The maximization of E in (2.2a) subject to constraints (2.4) is a linear programming problem. The critical line algorithm starts by using George Dantzig's "simplex algorithm" to maximize E. In addition to supplying the E-maximizing portfolio, the simplex algorithm renders certain other services to the critical line algorithm.

The simplex algorithm has two phases: phase 1 finds a feasible portfolio; phase 2 finds an optimum portfolio. Suppose that the equations in (2.4a) are written so that $b \geqslant 0$. Phase 1 finds X and X^D to minimize

$$z = \sum_{i=n+1}^{n+m} X_i \tag{5.8a}$$

subject to

$$AX + IX^D = b, \tag{5.8b}$$

$$X \geqslant 0, \qquad X^D \geqslant 0 \tag{5.8c}$$

where

$$X_D = \begin{pmatrix} X_{n+1} \\ \vdots \\ X_{n+m} \end{pmatrix} \tag{5.8d}$$

is an m component vector of "dummy slack" variables. If the least value of z is positive, then the original problem is infeasible. If an X, X^D is reached with $z = 0$ then X is a feasible solution to the original problem and the start of phase 2. Thus, one service which the simplex algorithm provides the critical line algorithm is to determine whether or not the model is feasible.

The minimization of z in phase 1 uses the fact that if a z-minimizing solution X, X^D exists then a *basic* optimum solution exists. Using terminology borrowed from our discussion of critical lines, a basic solution to the phase 1 linear programming problem designates m variables as IN (called "a basis" in usual linear programming parlance) and the remaining variables as OUT. It also requires A_{IN} to be nonsingular for the IN-set to be deemed a basis. The basic solution for the given IN-set is

$$X_i = 0 \quad \text{for } i \in \text{OUT}, \tag{5.9a}$$

$$A_{\text{IN}} X_{\text{IN}} = b, \tag{5.9b}$$

i.e.,

$$X_{\text{IN}} = A_{\text{IN}}^{-1} b. \tag{5.9c}$$

If $X_{\text{IN}} \geqslant 0$ in (5.9c) then it is a *feasible* basic solution.

The phase 1 calculation starts with the basic feasible solution $X^D = b$; feasible, that is, for the phase 1 problem. In a series of iterations the simplex algorithm produces a sequence of basic feasible (for the phase 1 problem) solutions with nonincreasing z. It is guaranteed that in a finite number of steps an optimal basic solution to the phase 1

problem is reached, and is recognized as being optimum. If phase 1 succeeds in reducing z to zero, then the last basic solution of phase 1 becomes the first basic solution of phase 2.

Phase 2 then iterates similarly, producing a sequence of basic, feasible solutions with nondecreasing phase 2 payoff (i.e., increasing E in our case). Phase 2 terminates when it either determines that payoff (E) is unbounded, or provides an E maximizing basic solution. Thus the principal service which the simplex algorithm renders the critical line algorithm is to determine whether E is bounded and, if so, what is an E-maximizing basic IN-set.

Below we treat the case in which there is one and only one E-maximizing solution. We note here, however, that if E is unbounded there is still a piecewise linear complete, nonredundant set of efficient portfolios. One of its segments is unbounded (in the direction of increasing E).

One fine point is important if we are to solve the portfolio selection problem for all possible input: μ, A, b, C. If A is rank deficient it is impossible to find a nonsingular subset A_{IN}. In this case, the optimum basic solution to the phase 1 problem will include one or more dummy slack variables X_i, $i \in [n+1, n+m]$, in the IN-set, though their values will equal zero if the model is feasible. Such dummy slack variables are left in the model but constrained to be zero, i.e., one adds

$$X_i \leqslant 0 \quad \text{for } i \in [n+1, 2n], \tag{5.10a}$$

$$z \equiv X_{n+m+1} = \sum_{i=n+1}^{n+m} X_i$$

as well as generally $X_i \geqslant 0$ as constraints. Alternately, add

$$z \leqslant 0. \tag{5.10b}$$

The model, thus augmented, has the same solution as the original model and is not rank deficient.

Thus, if A is rank deficient, another service which the simplex algorithm renders is to supply an equivalent model which is not rank deficient. This helps us to solve the portfolio selection problem for any input.

An upperbound on a variable such as in (5.10a), (5.10b) or, more generally,

$$X_i \leqslant u_i \tag{5.11a}$$

can be incorporated into a model of form (2.4) by adding an equation and a slack variable. In a subsequent section we sketch a computationally much more efficient procedure for handling upper bounds such as (5.11a) in general, and (5.10a) or (5.10b) in particular. It can also handle lower bounds other than zero:

$$X_i \geqslant l_i. \tag{5.11b}$$

Now let us see how the optimum basic solution to the E-maximization problem determined by the simplex algorithm determines the first critical line for the critical line algorithm. The first critical line uses the same IN-set as the optimum basic E-maximizing solution. Since $|A_{\text{IN}}| \neq 0$

$$|M_{\text{IN}}| = \begin{vmatrix} C_{\text{IN}} & A'_{\text{IN}} \\ A_{\text{IN}} & 0 \end{vmatrix} = -|A_{\text{IN}}|^2 \neq 0.$$

It is shown that on the first critical line

$$\begin{pmatrix} X_{\text{IN}} \\ \lambda \end{pmatrix} = M_{\text{IN}}^{-1} \begin{pmatrix} 0 \\ b \end{pmatrix} + M_{\text{IN}}^{-1} \begin{pmatrix} \mu_{\text{IN}} \\ 0 \end{pmatrix} \lambda_E,$$

X_{IN} does not vary—for all λ_E it equals the E-maximizing portfolio—only λ and λ_E vary. It is also shown that for some $\lambda_E^{\text{LOW}} \geqslant 0$

$$\eta \geqslant 0 \quad \text{for all } \lambda_E \geqslant \lambda_E^{\text{LOW}}.$$

Thus, according to (5.6), X_{IN} is an efficient portfolio. If $\lambda_E^{\text{LOW}} = 0$ then X_{IN} also has minimum V and provides the only efficient EV combination. Otherwise some $\eta_i \downarrow 0$ before $\lambda_E \downarrow 0$, and the critical line algorithm proceeds as described previously.

5.4. The critical line algorithm with upper bounds

In Section 5.3 we asserted that upperbounds on variables (5.11a) can be introduced into the critical line algorithm in a computationally more efficient manner than introducing an additional equation and slack variable into (2.4). In this section we sketch how this is done.

Each iteration of the critical line algorithm described in Section 5.2 partitions variables into two sets: IN and OUT. The portfolios traced out by the algorithm are known to be efficient because they satisfy the Kuhn–Tucker conditions (5.6). For a given IN/OUT partition the portfolio X varies linearly with λ_E in accordance with (5.4) and (5.5a), and so does η per (5.5b). The line segment traced out by varying λ_E remains efficient until one of the events in (5.7) occurs.

The algorithm modified to handle upper bounds (5.11a) is a slight variation of the above. The modification can also handle lower bounds other than zero as in (5.11b). The modified algorithm partitions variables into three sets: UP, DOWN and IN. The DOWN variables are fixed at their lower bounds; the UP variables are fixed at their upper bounds; and the IN variables are free to vary on the critical line. The portfolios generated by the modified algorithm are known to be efficient because they satisfy Kuhn–Tucker conditions

$$l_i \leqslant X_i \leqslant u_i \quad \text{for all } i, \tag{5.6a′}$$

$$\eta_i = 0 \quad \text{for } i \in \text{IN}, \tag{5.6b′}$$

$$\eta_i \geqslant 0 \quad \text{for } i \in \text{DOWN}, \tag{5.6c′}$$

$$\eta_i \leqslant 0 \quad \text{for } i \in \text{UP}. \tag{5.6d′}$$

The right-hand side of Equation (5.4) has to be modified to reflect the fixed values of X_i in UP and DOWN, namely

$$M_{\text{IN}} \begin{pmatrix} X_{\text{IN}} \\ \lambda \end{pmatrix} = \begin{pmatrix} 0_{\text{IN}} \\ b \end{pmatrix} - \widetilde{M}_{\text{IN}}(X_{\text{UP}} + X_{\text{DOWN}}) + \begin{pmatrix} \mu_{\text{IN}} \\ 0 \end{pmatrix} \lambda_E \tag{5.4′}$$

where $\widetilde{M}_{\text{IN}}$ contains only the IN rows but all the columns of M; X_{UP} is a vector whose ith component equals u_i for $i \in$ UP and zero otherwise; and X_{DOWN} is a vector whose ith component equals l_i for $i \in$ DOWN and zero otherwise. The events which signal ends of efficient segments are

$$X_i \uparrow u_i, \quad i \text{ IN}, \tag{5.7a′}$$

$$X_i \downarrow l_i, \quad i \text{ IN}, \tag{5.7b′}$$

$$\eta_i \uparrow 0, \quad i \text{ UP}, \tag{5.7c′}$$

$$\eta_i \downarrow 0, \quad i \text{ DOWN}, \tag{5.7d′}$$

$$\lambda_E \downarrow 0. \tag{5.7e′}$$

This modification of the critical line algorithm to handle upper bounds and nonzero lower bounds is an obvious adaptation of the Dantzig (1954) modification of the simplex algorithm for the same purpose.

5.5. The critical line algorithm with factor and scenario models of covariance

In Section 3.2 we referred to "computational advantages" which could be obtained if any of certain models of covariance are adopted. In this section we will illustrate such computational advantages with the one-factor model of covariance, and sketch the same for the many-factor, the scenario, the mixed scenario and factor models and for historical covariance matrices when there are more securities than observations. In problems with a large number of securities, computation time may differ by orders of magnitude between using a dense covariance matrix (e.g., as generated by equations such as (3.5)) and using the diagonal or nearly diagonal covariance matrices permitted by the aforementioned models of covariance.

Following the discussion in Section 3.2 we write portfolio variance as

$$V_p = \sum_{i=1}^{n+1} X_i^2 V_i \tag{5.12a}$$

where we write

V_i for $V(u_i)$, $i = 1, \ldots, n$,
V_{n+1} for $V(F)$,
X_{n+1} for B.

For this example we assume the standard constraint set (1.2a), (1.2b).

In addition, we need a constraint that defines X_{n+1} as the beta of the portfolio

$$\sum_{i=1}^{n} \beta_i X_i - X_{n+1} = 0. \tag{5.12b}$$

Recall that X_{n+1} does not enter the budget constraint (1.2a). Also, it is not required to be nonnegative in (1.2b); we therefore may treat it as always IN. Portfolio expected return remains as in (2.2a), with $\mu_{n+1} = 0$. The Lagrangian (5.1) then may be written as

$$L = 0.5 \sum_{i=1}^{n+1} V_i X_1^2 + \lambda_A \left(\sum_{i=1}^{n} X_i \right) + \lambda_B \left(\sum_{i=1}^{n} \beta_i X_i - X_{n+1} \right) - \lambda_E \left(\sum_{i=1}^{n} \mu_i X_i \right) \tag{5.13}$$

Taking partial derivatives we have

$$\eta_i = V_i X_i + \lambda_A + \lambda_B \beta_i - \lambda_E \mu_i, \quad i = 1, \ldots, n, \tag{5.14a}$$

$$\eta_{n+1} = V_{n+1} X_{n+1} - \lambda_B. \tag{5.14b}$$

Since X_{n+1} is always IN we have

$$\lambda_B = V_{n+1} X_{n+1} \tag{5.15a}$$

for all efficient portfolios. From (5.14) and (5.6b), on a given efficient segment we have

$$X_i = -\lambda_A (1/V_i) - V_{n+1} (\beta_i / V_i) X_{n+1} + \lambda_E (\mu_i / V_i), \quad i \in [1, n] \cap \text{IN}. \tag{5.15b}$$

For i IN and $\leqslant n$, substitute (5.15b) into (1.2a) and (5.12b) to obtain:

$$\begin{aligned} a\lambda_A + bX_{n+1} &= 1 + c\lambda_E, \\ d\lambda_A + eX_{n+1} &= f\lambda_E \end{aligned} \tag{5.16}$$

with constants a, b, c, d, e, f and variables λ_A, X_{n+1}, λ_E. Think of this as two equations in two unknowns, λ_A and X_{n+1}, and with λ_E as a parameter. Solve for λ_A and X_{n+1} as linear functions of λ_E; substitute this for λ_A and X_{n+1} in (5.15b) for all $i \in [1, n] \cap \text{IN}$. This expresses X_i as a linear function of λ_E on the critical line; i.e., we now have α_{IN} and β_{IN} in (5.5a). Substitute into (5.14a) to obtain γ_{IN}, δ_{IN} in (5.5b). We can now proceed as before to determine which event in (5.7) occurs first—or in (5.7′) if we treat upper bounds as in Section 5.4—to determine the new IN set, and repeat the process until $\lambda_E \downarrow 0$.

Note that the computations in this section require $O(n_{\text{IN}})$ calculations per iteration rather than the $O(n_{\text{IN}}^2)$ required to update and apply M_{IN}^{-1} in the general critical line computation of Section 5.1.

Analogous procedures can be developed if X is subject to the general constraint set (2.4a), (2.4b) and if the model of covariance assumes several factors; that is, if the return

on the ith security is

$$r_i = \alpha_i + \sum_{k=1}^{K} \beta_{ik} F_k + u_i, \quad i = 1, \ldots, n, \tag{5.17}$$

where F_k is the kth common factor, and u_i is an idiosyncratic term assumed uncorrelated with F_k, $k = 1, \ldots, K$, and all u_j with $i \neq j$. In the one-factor model we reduced the problem to solving two equations in two unknowns. When there are m constraint equations in (2.4a) and K factors in (5.17) the analogous procedure reduces the problem to solving $m + K$ equations for m Lagrangian multipliers and K portfolio exposures to the K factors. These $m + K$ variables are expressed as linear functions of λ_E, are substituted into n equations analogous to (5.15b) which thereby provide X as a linear function of λ_E, and the critical line algorithm proceeds as described previously. When m and K are much smaller than n, as is typical, this procedure reduces computation time manyfold as compared to updating and applying M_{IN}^{-1}.

A scenario model analyzed by Markowitz and Perold (1981a, 1981b) assumes that one of S mutually exclusive scenarios will occur with probability P_s, $s = 1, \ldots, S$. If scenario s occurs then the return on the ith security is

$$r_i = \mu_{is} + u_{is} \tag{5.18a}$$

where $E(u_{is}) = \text{cov}(u_{is}, u_{js}) = 0$ for $i \neq j$. Let $V_{is} = E(u_{is}^2 | s)$. The expected return E of the portfolio is still given by (2.2a) with

$$\mu_i = \sum_{s=1}^{S} P_s \mu_{is}. \tag{5.18b}$$

The μ_i can be computed from the above in advance of the optimization calculation. Let

$$X_{n+s} = \sum_{i=1}^{n} X_i (\mu_{is} - \mu_i), \quad s = 1, \ldots, S, \tag{5.18c}$$

be the expected value E of the portfolio, given that scenario s occurs, less portfolio grand mean E. The variance of the portfolio is

$$V_p = E(R_p - E_p)^2 = \sum_{s=1}^{S} P_s E(R_p - E_s + E_s - E)^2 = \sum_{i=1}^{n+S} X_i^2 \widetilde{V}_i \tag{5.19}$$

where

$$\widetilde{V}_i = \sum_{s=1}^{S} P_s V_{is}, \quad i = 1, \ldots, n,$$
$$\widetilde{V}_{n+s} = P_s, \quad s = 1, \ldots, S.$$

Thus V_p can be expressed as a positively weighted sum of squares in the n original variables and S new variables that are linearly related to the original variables by (5.18c).

We can now proceed as in the case of an S factor model. Markowitz and Perold (1981a, 1981b) also treat models that combine both scenarios and factors.

Markowitz et al. (1993a, 1993b) treat the case in which T historical periods (e.g., months or days) are used to estimate covariances among n securities. Let

$$X_{n+t} = \sum_{i=1}^{n} X_i(r_{it} - \mu_i), \quad t = 1, \ldots, T, \tag{5.20}$$

where r_{it} is the return on the ith security during period t. X_{n+t} then is the difference between portfolio return in the tth period and portfolio expected return E_p. Therefore

$$TV_p = \sum_{t=1}^{T} X_{n+t}^2. \tag{5.21}$$

Thus V_p is proportional to a sum of squares in new variables which are linearly related to the old. This is useful if it materially reduces the number of nonzero entries in M; that is, if T is much less than n.

Finally, Jacobs, Levy and Markowitz (2005, 2006) treat cases with factor, scenario and historical models of covariance when both short and long positions are allowed. In order to represent real-world constraints on portfolio choice, it is convenient for long/short models to have $2n$ variables (plus perhaps cash, slack and dummy variables) where the first n variables represent long positions and the second n represent short positions in the same stocks. If covariances among the long positions satisfy the assumptions of, e.g., a factor model, the $2n$-variable long/short model will not, since the "idiosyncratic" terms, the u_i in (3.4), now for $i = 1, \ldots, 2n$, are no longer uncorrelated with each other. Jacobs et al. provide a broadly applicable sufficient condition, "trimability", under which an existing factor, scenario or historical-covariance optimization program can be used, as is, to solve the long/short portfolio selection problem. Jacobs et al. call a model trimable if, for any feasible portfolio which holds a long and short position in the same security, it is feasible to reduce these two positions without changing the holdings of any other *risky* security and without reducing expected return. This condition assures that a simultaneous short and long position in the same security is not required to achieve any efficient EV combination. This, in turn, implies that existing factor and scenario model programs which "mistakenly" ignore the negative covariance among certain idiosyncratic risks, nevertheless produce a correct solution.

5.6. *Notes on computing*

Computing times become quickly obsolete. Speaking generally, by the beginning of the Twenty-First Century, efficient frontiers for a few thousand securities could be generated in seconds on a fast PC when a factor or scenario model, with a moderate number of factors or scenarios, was used.

Time required per iteration is roughly proportional to volume of nonzero entries in the M matrix. The number of iterations (corner portfolios) usually does not exceed,

or exceed by much, the number of securities in the universe. Often a large number of inconsequential iterations can be avoided by stopping the calculation when $\lambda_E \downarrow \varepsilon > 0$ for some small ε, e.g., $\varepsilon = 0.0001$, rather than continuing to $\lambda = 0.0$. The neglected iterations reduce V a negligible amount.

One circumstance when computing time may be important is when testing policies for estimating input parameters. A backtest or Monte Carlo run may require many re-optimizations, and an experiment may backtest or simulate a large number of policy combinations and/or constraint settings. For such experiments, the time saved by using factor or analogous models and generating frontiers only down to $\lambda_E \downarrow \varepsilon$ rather than $\lambda_E \downarrow 0$ may be appreciable.

The problem of minimizing V for a fixed E subject to constraints (2.4), or of minimizing $V/2 - \lambda_E E$ for fixed λ_E subject to (2.4), are quadratic programming problems. Quadratic programming, in turn, is a special case of convex nonlinear programming. (See, e.g., Fletcher (1987).) Thus, one might find it handy to compute specific points on the mean-variance frontier using an available quadratic, convex or nonlinear programming code such as solver in Excel. The critical line algorithm generates all points on the mean-variance frontier in the time required to compute a few points using a good quadratic programming code.

6. Separation theorems

6.1. The Tobin–Sharpe separation theorems

The Tobin (1958) model assumes that investors seek mean-variance efficiency subject to the standard constraint set (1.2a), (1.2b) where one security has zero variance, and the remainder have a nonsingular covariance matrix. Here it will be convenient to let a portfolio have $n+1$ components $X' = (X_1, \ldots, X_{n+1})$ where X_{n+1} is the zero variance ("risk-free") security. Then (1.2a), (1.2b) become

$$\sum_{i=1}^{n+1} X_i = 1, \tag{6.1a}$$

$$X_i \geqslant 0, \quad i = 1, \ldots, n+1. \tag{6.1b}$$

We will sometimes refer to μ_{n+1} as r_f, the return on the risk-free asset.

Tobin shows that X_{n+1} is IN on only one efficient segment. If r_f were greater than all the μ_i, $i = 1, \ldots, n$, then $X_{n+1} = 1$, $X_i = 0$, $i = 1 \ldots, n$, would be the only efficient portfolio. Assume that $r_f < \max \mu_i$. Tracing out the efficient set from high to low E, the efficient set starts with a portfolio with maximum E (i.e., the portfolio containing only the security with maximum μ_i when this is unique) and proceeds in the direction of decreasing E, σ and λ_E as discussed in the preceding section. At some point X_{n+1} must enter the IN set, since the portfolio with $X_{n+1} = 1$ is the efficient portfolio with minimum V. Let $X = X^T$ (T, for tangency) be the mix of risky securities at the

corner portfolio where X_{n+1} enters the IN set. Then the final (lowest E and V) efficient segment has

$$X^* = \left(X_1^T, \ldots, X_n^T, 0\right)' \tag{6.2a}$$

at one end and

$$X_r = (0, \ldots, 1)' \tag{6.2b}$$

at the other.

This may be shown along the following lines: when cash is IN the requirement that $\eta_{n+1} = 0$ implies

$$\lambda_1 = r_f \lambda_E$$

since $\sigma_{i,n+1} = 0$ for $i = 1, \ldots, n+1$. Substituting this into the Tobin model special case of (5.4), we find that the IN variables (other than X_{n+1}) are proportional to λ_E. In turn, substituting this into the Tobin model version of (5.2) we find that the η are also proportional to λ_E. Thus no IN $X_i \downarrow 0$ or OUT $\eta_i \downarrow 0$ until they all reach zero simultaneously when $\lambda_E = 0$, signaling that σ_{MIN} has been reached.

The Tobin theorem is illustrated in Figure 3. Let cTa be the set of efficient $E\sigma$ combinations of a portfolio analysis containing n risky securities subject to a standard constraint set (1.2a), (1.2b). It follows from (4.5), and the assumption that the covariance

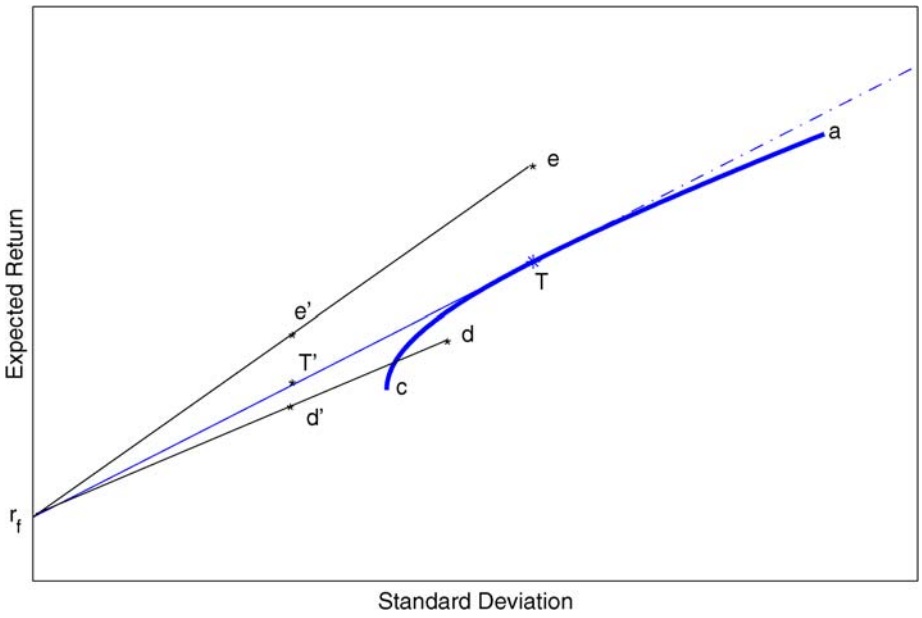


Fig. 3. The Tobin–Sharpe tangent portfolio.

matrix for the risky securities is nonsingular, that the efficient $E\sigma$ for the augmented, $n + 1$ security, model is piecewise hyperbolic until X_{n+1} enters, then is linear. One can construct this curve from *cTa* as follows. Plot r_f on the vertical (E) axis. Draw a line through $(0, r_f)$ tangent to the *cTa* curve. Typically, the tangency occurs at a point where $\mathrm{d}E/\mathrm{d}\sigma$ on the efficient set is the same as the slope of the line through $(0, r_f)$. (It is also possible for the "tangency" to happen at the E-maximizing point a with the slope of the line through $(0, r_f)$ less than or equal to the (left) derivative of the efficient frontier at a.)

The $E\sigma$ efficient set for the Tobin model is r_fTa—that is, it is the tangent line from $(\sigma, E) = (0, r_f)$ to the tangent point, then is identical to the efficient frontier without cash.

The Sharpe (1964) model drops the requirement that $X_{n+1} \geqslant 0$. Specifically, it assumes an investor constraint set consisting of (6.1a) and

$$X_i \geqslant 0, \quad i = 1, \ldots, n. \tag{6.1b$'$}$$

Negative X_{n+1} represents borrowing at the risk-free rate. For this model the set of efficient $E\sigma$ combinations is the entire ray in Figure 3, starting at $(0, r_f)$, passing through T and continuing without bounds in the direction of increasing E and σ. A point on this line with $E > E_T$ is a (σ, E) combination obtained by borrowing at the risk-free rate and leveraging the tangent portfolio T. The tangency (σ_T, E_T) combination is the efficient combination which maximizes the *Sharpe ratio*

$$S = \frac{E - r_f}{\sigma}. \tag{6.3}$$

The analytic argument for the Sharpe separation theorem is the same as that for the Tobin theorem, with the added observation that in the Tobin model the critical line with X_{n+1} IN is infeasible for $\lambda_E > \lambda_E^*$ where λ_E^* is the λ_E at which

$$X_{n+1} = 1 - \sum_{i+1}^{n} X_i \downarrow 0 \quad \text{as } \lambda_E \uparrow \lambda_E^*. \tag{6.4}$$

In the Sharpe model, $X_{n+1} \geqslant 0$ is not required. Since all the η_i and X_i other than X_{n+1} are proportional to λ_E on this critical line, none go to zero at any $\lambda_E > 0$. Therefore this critical line remains feasible and efficient for all $\lambda_E \geqslant 0$.

In presenting the Tobin model we noted that the tangency may occur at the point a, an undiversified portfolio. This may happen if μ and C are arbitrary; but cannot happen in the Sharpe–Lintner CAPM (Capital Asset Pricing Model). They assume that all investors have the same beliefs as well as being able to borrow all they want at the risk-free rate. It follows that all investors hold the tangent portfolio, plus perhaps borrowing and lending. Therefore security prices must adjust so that the tangent portfolio is the market portfolio therefore must contain all securities with positive prices. (See Markowitz (2005) for more on CAPM equilibrium when unlimited borrowing is not allowed; and Jacobs, Levy and Markowitz (2004) concerning the use of asynchronous simulation analysis for models too complex to solve with equilibrium analysis.)

6.2. Two-funds separation

In the Sharpe (1964), Lintner (1965), Mossin (1966) model, all efficient portfolios are linear combinations of the two portfolios:

$$X = X_{n+1}X_r + (1 - X_{n+1})X^* \tag{6.5}$$

for X_r, X^* in (6.2), and any $X_{n+1} \leqslant 1.0$. $X_{n+1} \in (0, 1]$ involves lending at the risk-free rate; $X_{n+1} < 0$ involves borrowing at that rate. Equation (6.5) is referred to as "*two-fund separation*".

Another case of two-fund separation is provided by the Roy model. We saw that the set of efficient portfolios for that model was a ray starting at the variance minimizing portfolio X_c and proceeding on a line LL' in the direction of increasing E. Let X_a be any point on this ray other than X_c. Then any point on the ray may be written as

$$X = yX_c + (1 - y)X_a \tag{6.6}$$

for y $\leqslant$ 1. A value of $y \in (0, 1)$ is interpreted as diversifying between X_c and X_a; whereas $y < 0$ is interpreted as shorting X_c—"shorting" in the sense of Roy where one may invest the proceeds of the short sale. Specifically, if $y < 0$, then $1 + |y|$ is invested in X_a. Equation (6.6) is thus an example of two-fund separation, since any efficient portfolio may be obtained by long or short positions in one or both of two portfolios.

6.3. Separation theorems not true in general

The Tobin and Sharpe results illustrated in Figure 3 may not hold if the set of efficient (σ, E) combinations, *cTa*, is derived subject to restrictions in addition to those of the standard model (2). For example, suppose that an efficient frontier for risky securities is derived subject to either of

$$X_1 \leqslant 0.05, \tag{6.7a}$$

$$X_2 \geqslant 0.01. \tag{6.7b}$$

First suppose that (6.7a) is required. It may be that the point e in Figure 3 would be feasible and would dominate T except that it is ruled out by requirement (6.7a). But, for sufficiently large $X_{n+1} \in (0, 1)$ the portfolio X in (6.5) will satisfy (6.7a) and will not be dominated by any mixture of cash and T. See e' and T' in the figure. T' is no longer part of the efficient set when cash is added.

Next suppose instead that (6.7b) is required. Then $(X^*, 0)$ may satisfy (6.7b), but X in (6.5) will not satisfy (6.7b) for sufficiently large $X_{n+1} \in (0, 1)$. Thus, again, some points on the straight line connecting $(0, r_f)$ and T may not be efficient, this time because they are not be feasible.

Inequalities (6.7a) and (6.7b) are inhomogeneous. Homogeneous equality and inequality constraints on $(X_1, \ldots, X_n)$ will continue to hold. For example, if

$$X_1 \leqslant 0.05 \sum_{i=1}^{n} X_i \tag{6.8}$$

holds for a portfolio X with $X_{n+1} = 0$ it will hold for X in (6.5) for all $X_{n+1} < 1.0$.

6.4. The Elton, Gruber, Padberg algorithm

Elton, Gruber and Padberg (1976) present simple criteria for determining the Tobin–Sharpe–Lintner tangent portfolio under certain special assumptions concerning the constraint set and covariance matrix. They consider two constraint sets. In both they assume that the investor can lend all it has or borrow all it wants at the risk-free rate; that is, they allow X_{n+1} to be chosen without a nonnegativity constraint in (3.1a). In the first of their constraint sets "short sales are allowed" in the sense of Roy; that is, they allow $X_i, i = 1, \ldots, n$, to be chosen without nonnegativity constraints. In their second model "short sales are not allowed". That is, they require

$$X_i \geqslant 0, \quad i = 1, \ldots, n. \tag{3.1b$'$}$$

One model of covariance they consider is the single-factor model of Sharpe (1963). The other they consider is the constant correlation model of Elton and Gruber (1973). This assumes that the correlation matrix (ρ_{ij}) satisfies

$$\rho_{ij} = \rho \quad \text{for all } i \neq j. \tag{6.9}$$

The constant correlation model is a special case of the single-factor model with $V(u_i)$ proportional to β_i^2 in (3.5). Elton, Gruber and Padberg (EGP) (1977) analyzes a model in which correlations are constant within industry groups, or sectors, but may vary between sectors.

EGP (1978) use the results of EGP (1976) to trace out the efficient frontier when there is no risk-free asset by, in effect, varying the r_f intercept in Figure 3. Markowitz (1999) conjectured that "Personally, I am not convinced that the EGP (1978) procedure will compute the efficient frontier faster than the critical line algorithm taking full advantage of the diagonalized versions of the one-factor or constant correlation models. The relative merits of the two algorithms on this problem remain to be determined". The conjecture is not correct. Jacobs et al. (2006) have since determined that Sharpe (1963) requires more computation per security per iteration than does EGP (1978).

Below we state, without proof, the EGP (1976) result for the one-factor model when "short sales are not allowed". This will illustrate EGP-type results.

To find the tangent portfolio for a given risk-free rate r_f, EGP (1976) maximize the Sharpe ratio S (6.3), subject to (1.2a), (1.2b). They distinguish three cases: (1) all $\beta_i \geqslant 0$, (2) all $\beta_i \leqslant 0$, (3) some $\beta_i > 0$ and some $\beta_i < 0$. In case (1), all $\beta_i \geqslant 0$, the Kuhn–Tucker conditions show that the n securities can be ranked from high to low according to the criteria

$$\frac{\mu_i - r_f}{\beta_i}. \tag{6.10}$$

If an i with a given $(\mu_i - r_f)/\beta_i$ is IN then all i higher on the list, with greater $(\mu_i - r_f)/\beta_i$, are also IN. Thus the IN set consists of the first k securities, $k \in [1, n]$, in

the list ordered by (6.10). It is further shown that, if k were known, the desired tangent portfolio could be computed as follows. For i IN, first compute

$$Z_i = \frac{\beta_i}{\sigma^2_{u(i)}}\left[\frac{\mu_i - r_f}{\beta_i} - \phi_k\right] \quad \text{if } \beta_i > 0, \tag{6.11a}$$

$$Z_i = \frac{\mu_i - r_f}{\sigma^2_{u(i)}} \quad \text{if } \beta_i = 0, \tag{6.11b}$$

where

$$\phi_k = \sigma^2_m \frac{\sum_{j=1}^{k} \frac{\mu_j - r_f}{\sigma^2_{\mu(j)}}\beta_j}{1 + \sigma^2_m \sum_{j=1}^{k} \frac{\beta_j^2}{\sigma^2_{\mu(j)}}}.$$

Portfolio weights X_i for IN i are then obtained by normalizing the Z_i:

$$X_i = Z_i \Big/ \sum_{j \leqslant k} Z_j \quad \text{for } i \leqslant k. \tag{6.12}$$

The cutoff k is determined so that $Z_k > 0$ and $Z_{k+1} \leqslant 0$. It is shown that this produces the S maximizing portfolio.

In case (2), all $\beta_i \leqslant 0$, a similar procedure applies except the securities are ranked from low to high $(\mu_i - r_f)/\beta_i$ rather than high to low. When stocks with both $\beta_i > 0$ and $\beta_i < 0$ are present EGP (1976) specify the following: follow the procedure outlined above to see which of the positive β stocks should be included in the portfolio. When no more positive β stocks are included, stocks with negative β should be tried, starting with the one with the smallest $(\mu_i - r_f)/\beta_i$, until no more enter. If any negative β stocks enter, it must decrease the size of ϕ_k, and so the highest excess return to positive β stocks previously rejected should be checked to see if any now enters. If more positive β stocks enter, then the negative β stock list should be checked and the procedure repeated iteratively until no more stocks enter.

6.5. *An alternate EGP-like algorithm*

Markowitz (1987), exercise 12.1, p. 288, and Christofi and Theodossiou (1995) present an algorithm for this model which is similar to EGP (1976) but without the need to go back and forth between the lists with positive and negative β_i. The Markowitz (1987) presentation minimizes

$$\frac{1}{2}V - E \tag{6.13}$$

subject to (6.1a) and (6.1b′). This will yield a portfolio Z somewhere on the linear efficient set of the Sharpe–Lintner model. Then Z can be normalized as in (6.12) to yield X.

Since X_{n+1} in (6.1a) represents lending (+) or borrowing (−), we will rewrite (5.12b) as

$$\sum_{i=1}^{n} \beta_i X_i - X_{n+2} = 0. \tag{5.12b′}$$

Here X_{n+2} represents portfolio beta. Use (6.1a) to eliminate X_{n+1} (and (6.1a) itself) from the problem. Letting

$$v_i = \mu_i - r_f \tag{6.14}$$

the Kuhn–Tucker conditions for minimizing (6.13) subject to (5.12b′) are

$$\text{if } \beta_i > 0, \quad X_i = \begin{cases} 0 & \text{for } \lambda \geqslant \hat{\lambda}_i = v_i/\beta_i, \\ (v_i - \lambda\beta_i)/V_i & \text{for } \lambda < \hat{\lambda}_i, \end{cases} \tag{6.15a}$$

$$\text{if } \beta_i < 0, \quad X_i = \begin{cases} 0 & \text{for } \lambda \leqslant \hat{\lambda}_i = v_i/\beta_i, \\ (v_i - \lambda\beta_i)/V_i & \text{for } \lambda > \hat{\lambda}_i, \end{cases} \tag{6.15b}$$

$$\text{if } \beta_i = 0, \quad X_i = \max(0, v_i/V_i) \quad \text{for all } \lambda \tag{6.15c}$$

where λ is the Lagrangian multiplier for the constraint (5.12b′). It follows that

$$f(\lambda) = \sum_{i=1}^{n} \beta_i X_i \tag{6.16}$$

is a *nonincreasing*, piecewise linear function of λ. According to (5.15a), the λ which solves the problem satisfies

$$V_{n+1} f(\lambda) = \lambda. \tag{6.17}$$

A unique solution to (6.17) exists, since its right-hand side rises strictly monotonically from $-\infty$ to $+\infty$ and its left-hand side is nonincreasing. The Markowitz (1987) algorithm sorts securities by v_i/β_i (like EGP). For convenience of notation, assume $v_1/\beta_1 \leqslant v_2/\beta_2, \ldots$. Use (6.15) to compute $f(\hat{\lambda}_1)$ and

$$b_0 = \sum_{\beta_i > 0} -\beta_i^2/V_i \tag{6.18}$$

the slope (zero or negative) of $f(\lambda)$ for all $\lambda \leqslant \lambda_1 = v_1/\beta_1$. This is used to determine if (6.17) occurs for $\lambda \in (-\infty, \hat{\lambda}_1]$. If not, in the general step for $k \geqslant 1$, having b_{k-1} and $f(\hat{\lambda}_k)$ in hand, determine b_k by adding or subtracting β_k^2/V_k from b_{k-1}, depending on whether $\beta_i < 0$ or $\beta_i > 0$; then compute

$$f(\hat{\lambda}_{k+1}) = f(\hat{\lambda}_k) + b_k(\hat{\lambda}_{k+1} - \hat{\lambda}_k). \tag{6.19}$$

If

$$V_{n+1} f(\hat{\lambda}_k) \leqslant \hat{\lambda}_k \tag{6.20}$$

then

$$f(\hat{\lambda}_k) + b_k(\lambda - \hat{\lambda}_k) = \lambda / V_{n+1} \tag{6.21}$$

solves (6.17). This is the case in which the desired λ is in $[\hat{\lambda}_k, \hat{\lambda}_{k+1}]$ for some k.

A minor variation handles the cases in which $\lambda > \hat{\lambda}_n$. The process takes no more computation than the EGP algorithm when no $\beta_i < 0$ and perhaps much less when both $\beta_i > 0$ and $\beta_i < 0$ are present.

The Christofi and Theodossiou algorithm uses a binary search for the λ that satisfies (6.17). The calculations of their algorithm are less directly comparable to EGP than are those of Markowitz (1987), but are most likely the fastest one when n is large.

7. Alternate risk measures

This section will sketch some of the properties and computational requirements of risk-return frontiers using risk measures other than standard deviation. Specifically we will consider semideviation, mean absolute deviation (MAD), expected shortfall, probability of loss, value-at-risk (VaR) and conditional value-at-risk (CVaR) as alternate measures. The question of how to judge the suitability to the investor's needs of the various measures is discussed in Section 8.

For each of these possible measures of risk, unless otherwise stated we assume that there is finite universe of security returns:

$$r_{it}, \quad i = 1, \ldots, n, \, t = 1, \ldots, T, \tag{7.1}$$

where n is the number of securities and T is the number of observations. These observations may be historical or a sample from a Monte Carlo model of joint returns.

7.1. Semideviation

Semivariance, proposed in Markowitz (1959), Chapter 9, differs from variance in that it only counts downside deviations from some level. Semideviation is the square root of semivariance. There are two versions of semivariance namely, semivariance from the mean

$$\mathrm{SV}_E = E\left\{(R_p - E)^-\right\}^2 \tag{7.2a}$$

and from some constant a:

$$\mathrm{SV}_a = E\left\{(R_p - a)^-\right\}^2, \tag{7.2b}$$

$$\mathrm{SD}_E = \sqrt{\mathrm{SV}_E}, \tag{7.2c}$$

$$\mathrm{SD}_a = \sqrt{\mathrm{SV}_a} \tag{7.2d}$$

where $c^- = -\min(0, c)$ is the negative part of c. We will refer to SV in statements which apply to either SV_E or SV_a.

Like mean-variance efficient sets, mean-SV efficient sets are piecewise linear in portfolio space, piecewise parabolic in (E, SV) space and consist of pieces which are usually hyperbolic but sometimes linear in (E, SD) space. For any portfolio X and observation t, the profitability of that portfolio at that observation (e.g., in that time period) is

$$\Pi(X, t) = \sum_{i=1}^{n} X_i r_{it}. \tag{7.3}$$

For a given portfolio X, we define P_X to be the set of t such that $\Pi(X, t) \geqslant E$ (or a), and define N_X to be the set of t such that $\Pi(X, t) \leqslant E$ (or a). $Z_X = P_X \cap N_X$. For a given portfolio X we define the semicovariance, SC_E or SC_a, between securities i and j as

$$\mathrm{SC}_E(i, j; X) = \sum_{t \in N_X} (r_{it} - \mu_i)(r_{jt} - \mu_j), \tag{7.4a}$$

$$\mathrm{SC}_a(i, j; X) = \sum_{t \in N_X} (r_{it} - a)(r_{jt} - a). \tag{7.4b}$$

Portfolio semivariance SV—either SV_E, or SV_a when $a = 0$ or (1.2a) holds—is related to the semicovariances SC (SC_E or SC_a respectively) as portfolio variance is to security covariances:

$$\mathrm{SV}_X = \sum\sum X_i X_j \mathrm{SC}(i, j; X). \tag{7.4c}$$

A segment of the efficient set in portfolio space is a portion of a critical line with the same formula (5.4) as in mean-variance analysis, except with semicovariances in place of covariances. The events which mark the ends of efficient segments include those which mark the ends of efficient segments in mean-variance analysis, namely those in (5.7), or in (5.7′) if upper bounds and lower bounds other than zero are included, plus two new types of events, namely,

$$\Pi(X, t) \downarrow E \text{ (or } a) \quad \text{for } t \in P_X \tag{7.5a}$$

and

$$\Pi(X, t) \uparrow E \text{ (or } a) \quad \text{for } t \in N_X. \tag{7.5b}$$

Then t moves from P_X to N_X in case (7.5a), or from N_X to P_X in case (7.5b). Markowitz (1959), Chapter 9 and Appendix A show how the M_{IN}^{-1} matrix is updated when the ($\mathrm{SC}(i, j; X)$) matrix is updated to reflect the new N_X set, and shows that the resultant critical line is a continuation of the set of efficient portfolios. At (7.5a) or (7.5b) $t \in Z_x$.

It is shown that, in case (7.5a) $t \in N_x - Z_x$, in case (7.5b) $t \in P_x - Z_x$, as the efficient set continues for smaller λ_E.

Markowitz et al. (1993b) show how to reduce computation time when the number T of observations is substantially less than the number of securities, under the assumptions

that imply (7.4c). Basic to their procedure is the observation that if we write

$$\sum_{i=1}^{n} r_{it} X_i - a = X_{tP} - X_{tN}, \quad t = 1, \ldots, T, \tag{7.6a}$$

or

$$\sum_{i=1}^{n} (r_{it} - \mu_i) X_i = X_{tP} - X_{tN}, \quad t = 1, \ldots, T, \tag{7.6b}$$

and require $X_{tP} \geqslant 0$, $X_{tN} \geqslant 0$ then minimizing SV for given E is equivalent to minimizing

$$\phi = \sum_{t} X_{tN}^2 \tag{7.7}$$

using (7.6a) for SV_a and (7.6b) for SV_E, since ϕ will be minimized with $X_{tP} X_{tN} = 0$. Thus by adding T new constraints and $2T$ new, fictitious securities we can express the risk objective as a sum of squares.

The Journal of Investing (1994) fall issue was devoted to a debate on mean-variance versus mean-semivariance. More recent work on semivariance is reported in Sortino and Satchel (2001). Also see the reference sections of the articles appearing in these two collections for the recent literature on semivariance, a.k.a. "downside risk" or "post-modern portfolio theory".

7.2. Mean absolute deviation (MAD)

An alternate measure of risk is mean absolute deviation

$$\mathrm{MAD} = E\big(|R_p - E_p|\big) \tag{7.8a}$$

where R_p is (as before) return on the portfolio, $E(\ )$ is the expected value operator and E_p is $E(R_p)$. It is recommended by Konno and Yamazaki (1991), and used in "stochastic programming" as proposed by Dantzig (1955) and discussed below.

An alternate measure is the mean absolute deviation from some constant:

$$\mathrm{MAD}_a = E\big(|R_p - a|\big). \tag{7.8b}$$

It might seem that yet another variation on MAD would be expected downside deviation

$$\mathrm{EDD} = E\big(|R_p - E|^-\big), \tag{7.8c}$$

$$\mathrm{EDD}_a = E\big(|R_p - a|^-\big). \tag{7.8d}$$

But since (with EUD_a suitably defined)

$$\mathrm{MAD}_a = \int_{-\infty}^{a} |R_p - a| \, \mathrm{d}P + \int_{a}^{\infty} |R_p - a| \, \mathrm{d}P = \mathrm{EDD}_a + \mathrm{EUD}_a, \tag{7.9a}$$

$$E - a = -\int_{-\infty}^{a} |R_p - a| \, \mathrm{d}P + \int_{a}^{\infty} |R_p - a| \, \mathrm{d}P = -\mathrm{EDD}_a + \mathrm{EUD}_a, \tag{7.9b}$$

it follows that

$$\mathrm{EDD}_a = \big(\mathrm{MAD}_a - (E - a)\big)/2. \tag{7.9c}$$

Thus the feasible portfolio which minimizes EDD_a for given E also minimizes MAD_a for that E. Letting $a = E$ shows that the same is true for EDD and MAD.

On the other hand, the value of a does make a difference. For example, if a portfolio A has a 0.5 probability of a 10% loss, but otherwise gains, and portfolio B has a 0.1 probability of a 20% loss but otherwise gains, then A has greater MAD_0 and B greater $\mathrm{MAD}_{-0.1}$.

Since

$$\mathrm{MAD} = (1/T) \sum_{t=1}^{T} (X_{iP} + X_{iN}) \tag{7.10}$$

for X_{iP} and X_{iN} in (7.6), the minimization of MAD for given E subject to constraints (2.4) is a linear programming problem; and thus the problem of finding (E, MAD) efficient sets is a parametric linear programming problem. See Dantzig (1963).

The MAD criteria has been used extensively in *stochastic programming* problems. In stochastic programming, a problem of risky action over time is represented as a linear programming problem in which some variables

$$X_{i1}, \quad i = 1, \ldots, n_1, \tag{7.11a}$$

represent actions taken "now"; while other variables

$$X_{itc}, \quad c = 1, \ldots, C_t, \, i = 1, \ldots, n_{tc}, \tag{7.11b}$$

represent actions taken at time t if contingency c occurs. The constraints on the choice of the Xs include a block of equations which constrain the choice of X_{i1}, and blocks which constrain X_{ict} for each c, t combination. The latter constraints may include variables with lower (earlier) t subscripts.

These systems may involve a huge number of variables and equations. See Ziemba and Mulvey (1998), and Volume 2 of this Handbook for a survey of applications; see Infanger (1994) for a discussion of how huge systems of this sort can be solved.

7.3. Probability of loss and value at risk (Gaussian R_p)

This section and the next consider two closely related possible measures of risk:

- the probability of loss or, more generally, the probability P_d of portfolio return less than d:

$$P_d = \mathrm{Prob}(R_p < d); \tag{7.12}$$

- the level of return, VaR, such that $\text{Prob}(R_p \leqslant -\text{VaR})$ is a given level P_g:

$$\text{VaR} = \{R_p \in (-\infty, \infty)\colon \text{Prob}(R_p \leqslant -\text{VaR}) = P_g\}; \tag{7.13}$$

e.g., the loss, $-\text{VaR}$, such that $\text{Prob}(R_p \leqslant -\text{VaR}) = 0.05$. This level is referred to as the "Value at Risk". See Jorion (1997). We first consider P_d and VaR when R_p is normally distributed, and then when they are to be estimated from a finite (historical or Monte Carlo) sample.

When R_p is normal, P_d is a function of portfolio mean E and standard deviation σ. Specifically,

$$P_d = N\left(\frac{d - E}{\sigma}\right) \tag{7.14}$$

where $N(\,)$ is the cumulative normal (0, 1) distribution. Roy (1952) recommended minimizing P_d. More generally, one could offer the investor an (E, P_d) efficient frontier from which to choose a portfolio. The portfolios which are (E, P_d) efficient are a subset of the (E, σ) portfolios, namely, all (E, σ) efficient portfolios whose E equals or exceeds that which maximizes $(E - d)/\sigma$.

Inverting (7.14) we see that, when R_p is normal, value at risk equals

$$\text{VaR} = b\sigma - E \tag{7.15a}$$

where

$$b = -N^{-1}(P_g) \tag{7.15b}$$

which is positive assuming $P_g < 0.5$. (7.15a) implies that every (E, VaR) efficient portfolio is EV efficient, but not vice versa. For an EV efficient point to be (E, VaR) efficient requires either

$$\left(\frac{\mathrm{d}\,\text{VaR}}{\mathrm{d}E}\right)_R = b\left(\frac{\mathrm{d}\sigma}{\mathrm{d}E}\right)_R - 1 > 0 \tag{7.16a}$$

or $(\mathrm{d}\,\text{VaR}/\mathrm{d}E)_R = 0$ and

$$\left(\frac{\mathrm{d}^2\,\text{VaR}}{\mathrm{d}E^2}\right)_R = b\left(\frac{\mathrm{d}^2\sigma}{\mathrm{d}E^2}\right)_R > 0 \tag{7.16b}$$

where $(\,)_R$ is the derivative to the right. A general analysis of (E, VaR) efficient sets when R_p is normal takes into account the fact that (σ, E) efficient curves may have kinks, and may have linear segments. The usual case of the latter is the segment which supplies smallest (σ, E) combinations when a risk-free asset exists (Figure 3). But linear segments of the (σ, E) efficient curve can also occur when a risk-free asset does not exist but a risk-free portfolio does (MT, 2000, Chapter 11).

7.4. Probability of loss and Value at Risk (non-Gaussian R_p)

From a computational point of view the most significant fact about P_d and VaR, when these are computed from a finite sample, is that a local minimum of either one is not necessarily a global minimum. This makes the derivation of (E, P_d) or (E, VaR) efficient sets exceedingly difficult when more than a few securities are involved. (See MacKay and Keefer, 1966, and Mausser and Rosen, 1999.)

The fact that a local minimum P_d is not necessarily a global minimum is illustrated in Figure 4. Recall that (as long as we do not have $\mu_1 = \mu_2 = \mu_3$) in a standard, three-security case the set of points (portfolios) which have the same mean is a straight line; e.g., the "iso-mean" line whose projection onto (X_1, X_2) is labeled $E = E_1$ in Figure 2(c). If E^* is strictly between the largest and smallest feasible E, then the iso-mean line $E = E^*$ will intersect the feasible set in a closed interval $[X^a, X^b]$ of positive length. Such an iso-mean line can be written as

$$X = \theta X^a + (1 - \theta) X^b. \tag{7.17}$$

The segment of (7.17) with $\theta \in [0, 1]$ is feasible. The θ for such an iso-mean line is plotted on the horizontal axis of Figure 4.

If $r_s = (r_1^s, r_2^s, r_3^s)$ is the return vector for sample observation s, then portfolio return for observation s is a linear function of θ:

$$R_p = r_s' X = \theta (r_s' X^a) + (1 - \theta)(r_s' X^b). \tag{7.18}$$

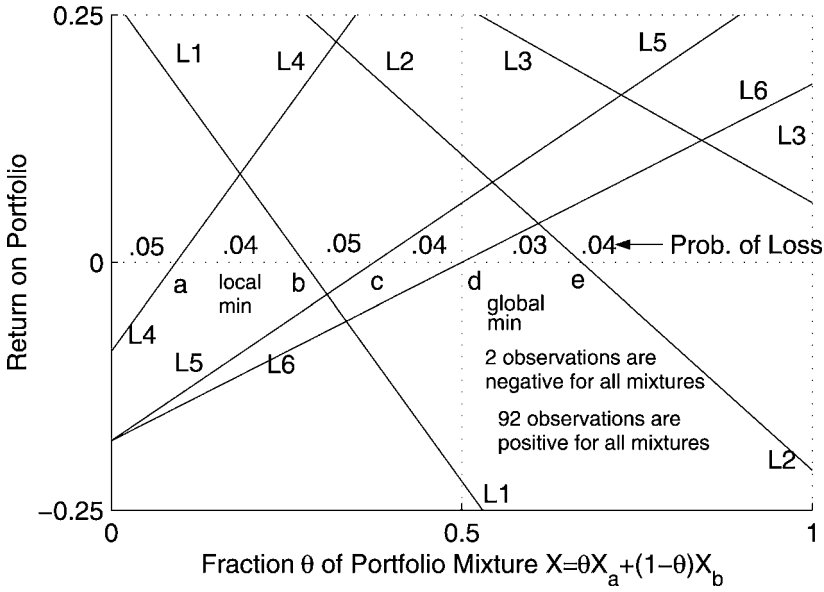


Fig. 4. Returns on mixtures of two portfolios for 6 of 100 observations, L1–L6.

In Figure 4 we illustrate a hypothetical case with sample size $S = 100$. For 92 observations $R_p > 0$ for all $\theta \in [0, 1]$ and are not shown in the figure; for 2 observations, $R_p < 0$ for all $\theta \in [0, 1]$ and are not shown. The R_p for six observations which intersect $R_p = 0$ for $\theta \in [0, 1]$ are shown, with R_p on the vertical axis. For example, the line labeled L1 shows portfolio profit on the vertical axis as a function of the fraction θ on the horizontal axis for observation 1, and similarly for lines L2 through L6.

For $d = 0$ in (7.12), at any point $\theta = \theta_0$ on the horizontal axis, P_d is the fraction of $R_p(\theta)$ lines below the θ-axis. In particular, in the interval $[0, a)$ $P_d = 0.05$ since 5 of the hundred lines have $R_p(\theta) < 0$. Similarly, P_d equals 0.04, 0.05, 0.04, 0.03, 0.04 respectively for $\theta \in [a, b]$, (b, c), $[c, d)$, $[d, e]$, $(e, 1]$. In particular, P_d has a local minimum at any point in $[a, b]$; but not a global minimum there.

Figure 5 uses the same lines L1–L6 to illustrate that a local minimum for VaR is not necessarily a global minimum. The broken line $\overline{abcdefg}$ is the $-$VaR curve for $p = 0.05$, since at any point on the curve there are 5 out of the 100 observations at or below the point on the line (therefore 5 below any point microscopically greater than the line).

Value at Risk is the negative of the magnitude plotted. $-$VaR has a local but not global maximum, therefore VaR a local but not global minimum, at d.

For 3 securities, minimum P_d or VaR can be computed by geometric means as illustrated in Figures 4 and 5, or their algebraic equivalent. With a few more securities the fact that a local optimum is not necessarily a global optimum makes tracing out a risk-return frontier a formidable task with these measures of risk.

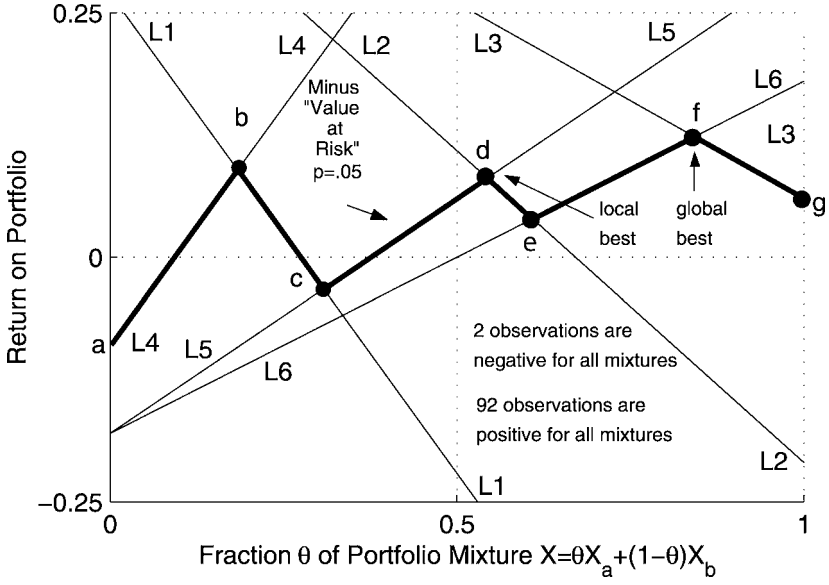


Fig. 5. Value at Risk for returns in Figure 4.

7.5. Conditional value at risk (CVaR)

Let VaR (for probability level P_g) be as in Equation (7.13). Let R_p^* be a random variable defined as follows

$$R_p^* = \begin{cases} R_p & \text{if } R_p \leqslant \text{VaR}, \\ 0 & \text{otherwise.} \end{cases} \tag{7.19}$$

Then conditional value at risk (CVaR) for level P_g is defined as

$$\text{CVaR} = E(R_p^*) = E(R_p \mid R_p \leqslant \text{VaR}), \tag{7.20}$$

i.e., CVaR is the "conditional expected return" given that $R_p \leqslant$ VaR.

CVaR does not suffer from the difficulty noted in Section 7.4 for VaR, that a local minimum is not necessarily a global minimum. See Rockafellar and Uryasev (2000) for properties of CVaR, a review of recent literature and a linear programming algorithm that can be used to determine (E, CVaR) efficient portfolios.

8. Choice of criteria

8.1. Exact conditions

If security returns $R = (r_1, \ldots, r_n)$ are joint normally distributed then the return R_p of the portfolio is normal. In this case the mean and variance are sufficient to determine the entire distribution of R_p. Tobin (1958) notes that mean and variance are sufficient, not only when R is normal, but also when the investor seeks to maximize the expected value of a concave *quadratic* single-period utility function. Pratt (1964) and Arrow (1971) argue that such a utility function is implausible because it implies that the investor's absolute risk aversion (i.e., aversion to gambles of a given dollar amount) increases with the wealth of the individual.

Normality is a sufficient but not a necessary assumption on the distribution of R for mean and variance to be sufficient for selecting portfolios. Specifically, Owen and Rabinovitch (1983) show that "elliptical distributions" also suffice. In terms of the notation of the present chapter, an elliptical distribution is defined as a joint distribution whose characteristic function is of the form

$$E_R\big(\exp\{\mathrm{i}t'R\}\big) \equiv \psi(t'Ct)\exp\{\mathrm{i}t'\mu\} \tag{8.1a}$$

for some function ψ, where $\mathrm{i} = \sqrt{-1}$, "exp" stands for exponentiation and t is an n-component parameter vector. (See, e.g., Feller (1970) for the definition of the characteristic function of a probability distribution.) When an elliptical distribution of R has a density function, it is of the form:

$$f(R) = c|C|^{-1/2}\xi\big((R-\mu)'C^{-1}(R-\mu)\big) \tag{8.1b}$$

for some function ξ. Owen and Rabinovitch assume C to be positive definite. Elliptical distributions may or may not have finite means and variances. When the latter two exist, μ and C in (8.1) are the mean vector and covariance matrix as elsewhere in this chapter. Special cases of elliptical distributions are the (multivariate) normal, the multivariate Student's t distribution and the symmetric distributions among the stable Paretian distributions.

When R is elliptical with finite means and variances, the risk averse investor chooses a mean-variance efficient portfolio. The efficient frontier is generated from μ, C and constraints as described previously. The only caveat is that one should not answer the question "what is the probability that $R_p \leqslant E - k\sigma$" by entries in a table of normal variates. See Owen and Rabinovitch (1983) and Ortobelli, Huber and Schwartz (2002) for elliptical distributions when finite covariances, and perhaps finite means, do not exist. Efficient sets are still piecewise linear, as in Sections 4 and 5, but C (and perhaps μ) are no longer covariances (and means).

Chamberlain (1983) analyzes a more general ("spherically symmetric") class of distributions for which the "Two Fund Separation" theorem holds when choice of portfolio is not subject to inequality constraints. The Owen and Rabinovitch results apply independently of the assumed constraint set. Ziemba (1974) provides an efficient method for computing approximate mean vs. dispersion efficient frontiers when returns have symmetric stable distributions.

8.2. *Mean-variance approximations to expected utility*

Chapters 6 and 13 of Markowitz (1959) illustrate empirically with a few utility functions and return distributions that, historically, average utility of return can be fairly closely approximated by

$$\mathrm{EU}(R) \cong U(0) + U'(0)E + 0.5U''(0)\left(V + E^2\right) \tag{8.2a}$$

or, usually better,

$$\mathrm{EU}(R) \cong U(E) + 0.5U''(E)V \tag{8.2b}$$

where $U(0)$, $U'(0)$ and $U''(0)$ represent the function and its first two derivatives evaluated at return $R_p = 0$ or, equivalently, wealth W equal current wealth. The right-hand side of (8.2a) is based on a quadratic Taylor approximation to U around the point $R = 0$; that in (8.2b), about $R = E$.

Young and Trent (1969) find similar results for a much larger sample of historical return distributions for $E[\log_e(1 + R)]$ or, equivalently, geometric mean GM equal to $\exp\{E[\log_e(1 + R)]\}$. That is, Young and Trent find that if you know the mean and variance of a historical return series you can guess fairly closely its average log.

Samuelson (1970) and Ohlson (1975) present conditions under which mean and variance are asymptotically sufficient for an optimum decision as the length of holding periods—i.e., the intervals between portfolio revisions—approaches zero. The Young

and Trent results indicate that, as far as $E[\log_e(1 + R)]$ is concerned, monthly holding periods are short enough to insure that $f(E, V)$ is virtually the same as $E(U)$.

Levy and Markowitz (1979) compare average utility versus mean-variance approximations for various utility functions, historical distributions, and methods of mean-variance approximation. The quadratic approximations analyzed by Levy and Markowitz depend on a parameter k, and are fit through the three points: $E - k\sigma$, E, and $E + k\sigma$. For the utility functions and historical distributions considered, the approximations improved as k approached zero. Since the right hand-side of (8.2b) is the limit of the Levy–Markowitz approximations as $k \to 0$, Levy and Markowitz concluded that (for the given sample and utility functions) (8.2b) is as good as any of their approximations. In our discussion of the Levy–Markowitz results, we quote results for $k = 0.01$, the smallest k they considered.

As in Markowitz (1959) and Young and Trent (1969), Levy and Markowitz (1979) assume that distributions of returns are the same as various historical distributions (or, at the least, have the same E, V, $E[\log_e(1 + R)]$, $E[(1 + R)^{1/2}]$, etc.). Distributions used include annual returns on 149 investment company portfolios, annual and monthly returns on 97 individual stocks, and annual returns on randomly drawn portfolios with $n = 5$ or 6. The utility functions analyzed were $\log_e(1 + R)$, $(1 + R)^a$ for $a = 0.1$, 0.3, 0.5, 0.7, and 0.9, and $-\exp[-b(1 + R)]$ for $b = 0.1$, 0.5, 1.0, 3.0, 5.0 and 10.0.

For $U = \log_e(1 + R)$ the correlation between $E(U)$ and the mean-variance approximation $f(E, V)$ over the 149 distributions of annual returns on investment companies, was $\rho = 0.997$. That for $V = (1 + R)^{1/2}$ was $\rho = 0.999$. For all utility functions except $-\exp[-b(1 + R)]$ for $b = 3.0$, 5.0, and 10.0, the correlation between $E(U)$ and $f(E, V)$ was at least 0.997. In most cases it was 0.999.

For $-\exp[-b(1 + R)]$, $b = 3.0$, 5.0, and 10.0, the correlations were respectively 0.949, 0.855, and 0.447. Concerning the "misbehaved" $-\exp[-10(1 + R)]$ Levy and Markowitz argue that while all utility functions satisfy the Von Neumann and Morgenstern (1944) axioms, not all will appeal to many (or any) investors. In particular,

$$U = -\exp\big[-10(1 + R)\big] \tag{8.3}$$

exhibits certain strange preferences. For example, an investor with this utility function would prefer 10 percent with certainty to a 50–50 chance of zero return (no gain, no loss) versus a blank check (e.g., a gain of 1 000 000 000 percent). Levy and Markowitz refer to this utility function as being pathologically risk averse.[5]

Correlations between $f(E, V)$ and $E(U)$ were smaller over the set of annual returns on 97 stocks. For example, for $U = \log_e(1+R)$ the correlation was 0.880 for the annual stock returns as compared with 0.997 for the annual investment company returns. (The mean-variance criteria is to be applied to the portfolio as a whole; single stocks were used as examples of "portfolios" with greater variability.) When the holding period was

[5] Markowitz, Reid and Tew (1994) refer to this as an implausibly low "value of a blank check" and present a survey suggesting a range of plausible values.

reduced the correlations increased, as expected. For example, for monthly stock returns the correlation for $U = \log_e(1 + R)$ was 0.995. Also, a bit of diversification helped a lot. For annual returns on 19 small (no overlapping) portfolios ($n = 5$ or 6) drawn at random from the 97 securities, the correlation for $U = \log_e(1 + R)$ was 0.998. Similar effects (of going from investment companies to stocks, from annual stock returns to monthly stock returns, and from annual stock returns to annual returns on 5- or 6-stock portfolios) were found for the other utility functions.

Thus, for most utility functions considered, the mean-variance approximations did quite well, especially for annual returns on diversified portfolios and for monthly returns (even on undiversified portfolios). The success of the mean-variance approximations leads to an apparent anomaly: for many utility functions, if you know mean and variance you practically know expected utility; the mean-variance approximation to expected utility is based on a quadratic approximation to the single-period utility function; yet Pratt and Arrow show that *any* quadratic utility function has the objectionable property of increasing absolute risk aversion. Levy and Markowitz (1979) show that the anomaly disappears if one distinguishes three types of quadratic approximations:

(1) Assuming that the investor has a utility-of-wealth function that remains constant through time—so that the investor moves along the curve to a new position as his or her wealth changes—fit a quadratic to this curve at some instant, and continue to use this same approximation subsequently.
(2) Fit the quadratic to the investor's current single-period utility function. For example, if the investor has an unchanging utility-of-wealth function, choose a quadratic to fit well near current wealth (i.e., near $R = 0$).
(3) Allow the quadratic approximation to vary from one portfolio to another; i.e., let the approximation depend on the mean, and perhaps the standard deviation, of the probability distribution whose expected value is to be estimated.

The Pratt–Arrow objection applies to an approximation of type (1). The quadratic approximation behind (8.2a) is of type (2); while that behind (8.2b) and the various Levy–Markowitz approximations are of type (3). Levy and Markowitz show that, under quite general assumptions, the type (3) mean-variance maximizer has the same risk aversion in the small as does the original expected utility maximizer; hence the Pratt–Arrow objection to "quadratic utility" does not apply.

More recent studies include Dexter, Yu and Ziemba (1980), Pulley (1981, 1983), Kallberg and Ziemba (1981, 1983, 1984), Kroll, Levy and Markowitz (1984) and Simaan (1993). Dexter et al. and Simaan each assume a particular (different in the two cases) functional form for the distribution of securities returns, and fit its parameters to historical data. Most of the other studies use historical returns as if they were the actual joint distributions. Pulley (1983) assumes particular functional forms, generates synthetic "histories" of joint returns from these forms, and then proceeds as if the samples were the joint distributions. Ederington (1986) evaluates mean-variance approximations in terms of "10 000 simulated years" by drawing four quarters each at random from quarterly returns on 130 mutual funds. Ederington argues that these 10 000 synthetic years for each mutual fund will provide more extreme cases, and therefore provide a

better test of mean-variance approximations, than the relatively few years of actual history used in the Young and Trent and the Levy and Markowitz analyses. Hlawitschka (1994) considers portfolios of puts and calls.

The conclusions of the above studies are generally supportive of mean-variance approximations. However, Grauer (1986) illustrates that if leverage is permitted mean-variance approximations may produce poor results unless the choice of portfolio is constrained to avoid bankruptcy.

8.3. Significance of MV approximations to EU

Von Neumann and Morgenstern (1944), Savage (1954) and their followers have convinced many that rational action in the face of risk or uncertainty involves maximizing the expected value of some utility function. The calculations cited in the preceding subsection indicate that, for certain single-period utility functions and historic return distributions, little is lost by using mean-variance approximations to expected utility. The disadvantage to using such approximations is that they do not provide the precisely optional solution. Their advantage is that they take less computer time, have fewer parameters to estimate and one does not have to ascertain the investor's utility function. Insofar as the approximation is satisfactory, by picking a suitable point off the MV efficient set, the investor can approximately maximize his or her expected utility without having made this utility function explicit.

Markowitz and van Dijk (2003) consider loss of utility when a mean-variance surrogate is used in place of the "derived" utility function (Bellman, 1957) in certain dynamic investment games with trading costs and changing probability distributions. They report little loss in expected utility as compared to an optimum strategy. The significance of these experiments is that the mean-variance surrogate can be scaled up to apply to much larger investment games than can current dynamic optimization techniques. In particular, the larger games have too many state variables for dynamic programming and too many time periods for stochastic programming.[6]

9. Risk-return analysis in practice

In this section we discuss issues involved in the practical application of risk-return analysis—issues such as choice of criteria, estimation of parameters, and the uncertainty of parameter estimates.

[6] When trading costs are zero, probability distributions constant and the investor objective to maximize the expected value of a logarithmic or power function of final wealth, Mossin (1968) and Samuelson (1969) have shown that the current period derived utility function is essentially the same as the final (power or logarithmic) utility function. Thus the observations in Section 8.2, on the efficacy of mean-variance approximations, apply immediately to question of choice of portfolio in these T-period games. See also Hakansson (1971) for the case with serial correlation.

9.1. Choice of criteria

As already noted, opinions differ as to choice of criteria for a risk-return analysis. The following subsections assume that the decision has been made to use mean-variance as criteria. Here we provide some reasons that might incline a management team toward these criteria. We assume that the investor is a risk-averse institution such as a pension or endowment fund.

The tests of the ability of functions of mean and variance to approximate expected utility, reported in Section 8, provide theoretical justification for the use of mean and variance as risk-return criteria. For the most part, similar tests are lacking for other criteria.

An examination of the approximating utility functions implied by mean-variance, mean-"expected loss", mean-"mean absolute deviation" and mean-"probability of loss" seem to us to favor the first of these. See Markowitz (1959), Chapter 13 for details. In particular, in choosing among probability distributions with the same gains but different loss probabilities, the last three criteria are not risk averse to changes in probabilities of loss. E.g., the expected loss criteria would be indifferent between a 50–50 chance of a 10% versus a 30% loss rather than have a 20% loss with certainty. Variance as criteria would prefer the certainty to the chance with equal expected value. Similar comments apply to mean-absolute-deviation as a measure of risk, with downside deviation from the mean substituted for actual loss in the analysis. Our choice of mean-variance below is based, in part, on an assumed desire for risk-aversion among alternate loss distributions.

The choice of mean-variance rather than mean-semivariance in what follows is based on two considerations. The first is an empirical test not heretofore reported in the literature. Using the model in Bloch et al. (1993), Peter Todd tested how well the model did ex post—in terms of mean and semivariance—if the optimizer was told ex ante to minimize variance rather than semivariance. The result of this backtest was that the model did *almost* as well ex post, in terms of semivariance, when the optimizer was told ex ante to minimize variance as it did when it was told ex ante to minimize semivariance.

The second consideration is that if we knew the true joint distribution of security returns, then presumably semivariance would be the preferable measure. But we do not know this distribution. In subsequent subsections we consider alternate methods that have been proposed for estimating means, variances, and covariances. In particular, it has been reported that factor models do perceptibly better ex post than do historical covariances as estimates of future covariance. Also, various methods of estimating expected forthcoming return have been proposed, and some discussed below, as superior to using historical averages. Since it is more difficult to use such models and methods with semivariance than with variance, and Peter Todd's experiments suggest that it does not make much difference, we stick with mean-variance below.

It must be conceded, however, that it is not impossible to use such methods and models with semivariance. Risk-return analysis using the latter currently requires a sample of observations. Rather than using historically observed joint returns one could generate joint returns using a Monte Carlo model. The parameters of the Monte Carlo model

could, in turn, be based on the aforementioned methods and models. Further assumptions would be needed as to the form of the joint distribution; e.g., it could be assumed Gaussian, but in that case the mean-variance assumption would be sufficient. Since the following subsections of this section discuss the estimation of means, variances and covariances, but not further characteristics of the joint distribution, in effect we assume mean-variance as criteria.

It should also be conceded that the arguments above—in terms of the shapes of the approximating utility function associated with different risk measures—are not as convincing as when such arguments are combined with actual comparisons between approximation and expected utility. For example, Zenios and Kang (1993) examine the graphs of the approximating utility functions and decide that mean absolute deviation is acceptable. For another example, Markowitz (1959) argues that the maximum loss risk measure is inconsistent with expected utility, and therefore dismisses it as a risk measure for risk-return analysis. However, Young (1998) presents a short historical test (seven country stock indices, 30 in-sample, 30 out-of-sample months) and many Monte Carlo runs in which mean-variance and mean-maximum loss perform "similarly". If further comparisons confirm Young's results then it would be a matter of convenience as to which risk measure to use. By "convenience" we refer to convenience of parameter estimation as well as computational convenience. As noted, below we discuss the estimation of means, variances and covariances. These could be all of the inputs to a mean-variance analysis or, perhaps, part of the input to a Monte Carlo model which generates "observations" for a mean-semivariance, MAD or maximum loss analysis, e.g., as in Worzel, Vassiadou-Zeniou and Zenios (1994).

9.2. Tracking error or total variability

In the experiment conducted by Peter Todd, reported above, it made little difference, ex post, whether one told the optimizer ex ante to minimize variance or semivariance. On the other hand, it made a very large difference whether the model was told to minimize tracking error or minimize total variability (i.e., minimize variance without reference to some benchmark). If the optimizer was told, ex ante, to minimize tracking error [total variability] and then measured ex post in terms of total variability [tracking error] it did very much worse ex post than if it were given the correct objective in the first place. This is not surprising. For example, other things being equal it would be better to have a beta (of the portfolio against "the market" as benchmark) near zero if low total variability is sought, and near unity if low tracking error is sought.

The discussion in the following subsections is consistent with the choice of either tracking error or total variability as the measure of risk, since means, variances and covariances are the inputs required for either analysis when tracking error is defined, as usual, as the standard deviation of the difference between portfolio and benchmark return. (Tracking error can also be defined and minimized, for given expected return, in terms of MAD or maximum loss. See Worzel, Vassiadou-Zeniou and Zenios (1994).)

9.3. Estimates for asset classes

Often asset classes—such as large capitalization, small capitalization, foreign, emerging market, long term fixed income, etc.—are used as the "securities" of a mean-variance analysis. The latter is used to help decide a portfolio's allocation to such broad categories. This allocation is implemented either passively, by investing in index funds which track the respective asset classes, or actively by investing in (in-house or external) managers who are given the asset class indices as benchmarks to be tracked and outperformed, or by an active-passive strategy that does some of each.

The widespread use of this topdown analysis starting with an efficient set analysis at the asset class level is due, in great part, to the work of Brinson, Hood and Beebower (1986), Brinson, Singer and Beebower (1991), Brinson, Hood and Beebower (1995). See Swensen (2000) as an example of the use of mean-variance analysis at an asset class level by a large institutional investor with specific advice concerning the implementation of the asset allocation. See Gibson (2000) for its use by individual investors and their advisors.

Time series of returns for commonly used asset classes are available through http://www.ibbotson.com and http://www.russell.com. Over a century of returns on the equities of different countries is presented in Dimson, Marsh and Staunton (2002). If the purpose of a portfolio analysis is to guide current investment policy (rather than a theoretical or historical study) then its inputs should be estimates about the future rather than facts about the past. Historical statistics of asset class returns are often adjusted to reflect analysts' or management teams' views of the future.

9.4. Estimation of expected returns for individual equities

We know no procedure to put data in and get "correct" expected return estimates out. Our objective here is to briefly summarize, and cite sources for, some major alternatives that have been proposed for expected return estimates.

The *dividend discount model* asserts that the expected present value (EPV) of a stock is the present value of its expected future dividend payments. See Williams (1938) for the roots of the proposal, Gordon (1962) on the treatment of the dividend growth rate, and Farrell (1983, 1985) on estimation procedures. To use EPVs in a portfolio analysis, one can either define the μ vector to be EPVs, or can convert EPVs to expected returns. In the former case, the output of the portfolio analysis is an EPV vs. risk tradeoff curve. In the latter case, one needs to estimate the likely change of price during the period of analysis as a function of current price and EPV.

Any substantial in magnitude and statistically significant departure of average returns on classes of investments from their CAPM expected values is defined as an "anomaly". See, for example, Rosenberg, Reid and Lanstein (1984–85). This includes more (or less) than theoretically expected returns to size (capitalization) and other value measures. Anomalies also include observed "January effect", end of month effect, end of week

effect; a momentum effect, in some cases, and a reversion to the mean in others, and so on.

See Jacobs and Levy (1988) for an extensive survey of the previous anomaly literature. Before 1988 anomaly studies considered small numbers of variables, usually one to three at a time. Observing that some apparent anomalies may be surrogates for others, Jacobs and Levy fit a series of monthly cross-sectional regressions of security excess returns against 25 anomaly and 38 industry variables. This allowed them to "disentangle" what they called the "pure" (i.e., underlying) anomalous effects from what they called the "naive" effects observed from simple regressions against anomalous variables one at a time. The Jacobs and Levy methodology may be used for expected return estimation as well as for explaining observed anomalies. Further work along these lines include Haugen and Baker (1996) and Schwartz and Ziemba (2000).

Josef Lakonishok and various co-authors have investigated extensively the relationships between observables and expected returns. See, for example, Lakonishok, Shleifer and Vishny (1994) and Chan, Narasimhan and Lakonishok (1996). Examples from other important contributors to the expected return literature include Lamont (1998), Ohlson (1979) and Penman (1996).

Fama and French (1995) present a now widely used *three-factor model* which explains returns on individual stocks in terms of returns on indices of (a) return on "the market portfolio", (b) return on stocks with a high ratio of book value to market value, minus returns on stocks with a low such ratio (HML), and (c) returns on stocks with small capitalization minus that of stocks with big capitalization (SMB). If the Capital Asset Pricing Model (CAPM) of Sharpe (1964) and Lintner (1965) were precisely true, then the expected return of a security would depend only on its regression against the market portfolio. Fama and French find, to the contrary, that the expected return of a security does not depend on its sensitivity to moves in the general market but, rather, to its sensitivities to HML and SMB.

9.5. Black–Litterman

Following Savage (1954), Chapter 12 of Markowitz (1959) argues that the means, variances and covariances of a portfolio analysis should summarize the probability beliefs of an investor or investment management team. Chapter 2 of Markowitz (1959) includes some suggestions as to how these beliefs could be solicited. Black and Litterman (1991, 1992) argue that a great practical difficulty of this type of approach is that it requires estimates of expected returns for each security in the universe of analysis. This includes securities that the analyst or team has studied at length, and others about which they know little or nothing. Black and Litterman offer a procedure in which the analyst team can supply estimates for some securities (or asset classes) and not others. The procedure is such that if the analyst team expresses no beliefs then the process produces expected return estimates equal to CAPM equilibrium expected returns (for a given covariance matrix, estimated separately); if the analyst team supplies expected return estimates for all securities, and indicates strong beliefs for all of them, then the process almost re-

produces these as its expected return estimates; and, finally, if the analyst team supplies beliefs for some and not other securities the process produces expected return estimates between the preceding two extreme cases.

9.6. Security analyst recommendations

The first paragraph of Markowitz (1952), and Chapter 2 of Markowitz (1959), assume that the expected returns used as the inputs to portfolio analyses will be the outputs of prior security analyses. In fact, security analysts have been reluctant to estimate, e.g., the requisite expected returns for securities. One output of security analysts are earnings estimates. These are used in forward looking price-to-earnings ratios used in some value oriented formulas for estimating expected returns, discussed above. Another security analyst output is "buy", "sell" or "hold" recommendations. But portfolio theory implies that, even if a security analyst did a good job of estimating the expected return of a particular stock, he or she would not be able to determine whether it was advantageous to add, delete or hold the security in a portfolio. The decision to buy, sell or hold, and how much to buy if the security is to be purchased, depends not only on the expected return (and variance) of the security in question, but also on the expected return and variance opportunities presented by other securities, the covariances between the security in question and other securities, and the extent to which the various securities contribute to any constraints imposed on the choice of portfolio.

What could be done, however, would be to assign numbers, such as 1, 2, 3 or -1, 0, 1, to sell, hold and buy recommendations and use these as the expected returns of a portfolio analysis. The $E\sigma$ frontier, then, would show the average ranking of securities in the portfolio as a function of portfolio risk. See Barber et al. (2001) on the efficacy of security analyst recommendations.

9.7. Estimates of covariance

Sharpe (1963) reported that the one-factor model did better in an experiment than did a historical covariance matrix. Cohen and Pogue (1967) also reported favorably on the one-factor model as compared to more complex models. Often in such comparisons a historical covariance matrix is referred to as the "full Markowitz Model". In fact, Markowitz (1959), pp. 96–100, proposed a factor model rather than historical covariances. He did not, however, test the efficacy of alternative factor models as did Sharpe and others. If the market portfolio is used as the measure of the underlying factor then, by definition, the average beta (regression against the market) is 1.0. Blume (1975, 1979) reported a strong tendency for betas that deviate from this average to move towards unity, the average beta, over time.

Elton and Gruber (1973) found that a special case of the one-factor model, namely the "constant correlation model" did better than the more general one-factor model. The constant correlation model assumes that the correlation coefficient ρ_{ij} is the same for all $i \neq j$. It did better in the sense that efficient portfolios computed from estimates

of the constant correlation model did better ex post than those based on estimates of betas for the single-factor model which, in turn, did better than historic covariances. Since an arbitrary covariance matrix includes the one-factor model as a special case, and the latter includes the constant correlation model as a special case, it might seem curious that the special models outperform the general models. Apparently, historical deviations of covariances from the patterns implied by the simple models are more noise than actual structure.

King (1966) found that industry factors could be identified by a process of determining which pair of securities was most correlated; combining these two into a new fictitious security; repeating the process using the fictitious security instead of its two constituents; etc., at each step combining original or fictitious securities into new fictitious securities. Towards the end of this process, the remaining fictitious securities were easily identified as industry groups. A similar process might be of value currently to classify securities into asset classes.

If the one-factor model were correct then, when the market went up, a well diversified portfolio of high beta stocks would go up even faster. There turned out to be occasions when the market went up but high beta stocks went down, to an extent that could hardly be ascribed to chance. This contributed to the widespread interest in the Rosenberg (1974) model that includes industry and other factors. Of historical importance in the spread of portfolio theory in practice, Rosenberg founded BARRA which supplied factor models of covariance and educated practitioners in their possible uses.

The Rosenberg many-factor model, and the Fama–French three-factor model for that matter, relate co-movements among security returns to observable characteristics of the securities and their companies, such as their industry or size. Another approach is to look only at the historical covariance matrix of security returns and use some form of factor analysis to estimate the number of underlying factors which explain this observed covariance and the dependencies (betas) of each security on the factors. Such use of factor analysis for portfolio analysis dates back at least to Feeney and Hester (1967). Recent factor analysis of covariance is often referred to as an "APT factor model". The Ross (1976) Arbitrage Pricing Theory is actually a model of expected return in equilibrium, competitive with CAPM, and thus outside the scope of the present chapter. APT assumes, among other things, that returns are generated by a factor model and concludes that expected returns of securities are linearly related to their respective betas against the various factors. This conclusion of APT encouraged an interest, by APT adherents, in using factor analysis of covariance matrices to infer the factor structure of security returns. See Blin and Douglas (1987).

9.8. Parameter uncertainty

Even if "Nature" drew random returns from an unchanging joint distribution, observed means, variances and covariances would reflect random noise as well as underlying parameters. Citing studies of Jobson and Korkie (1981) and Frost and Savarino (1988), Michaud (1989, 1998) emphasize that portfolio optimization tends to magnify the ef-

fects of sampling error. For example, if past average returns are used as the expected return inputs to a mean-variance analysis, and if a particular security (or asset class) has done especially well during the historical sample period, in part at least because it was "lucky" or had special circumstances on its side, the optimizer may heavily overweight this investment. Similarly, the optimizer will be improperly attracted to pairs of securities which randomly had lower covariance.

In practice, unadjusted historical averages are less frequently used as estimates of forthcoming expected returns than are unadjusted historical covariance matrices as estimates of forthcoming covariances. (See Kallberg and Ziemba (1981, 1983) and Chopra and Ziemba (1993) concerning the effects of estimation errors: especially the much greater effects of errors in expected returns as opposed to errors in variance and covariances.) But whatever method of estimation is used—historical means and covariances, "judgment", or the outputs of some model—the inputs to a portfolio analysis will have estimation error. A common device to reduce the consequences of estimation errors is to impose constraints on the selected portfolios; e.g., it is common to impose bounds on the amount invested in each security or asset class, or on groups of them. Michaud objects that such bounds are ad hoc. They need not be completely so. Backtests are sometimes used to see what policy concerning bounds would have worked well in the past. Also, see Markowitz (2000) for a theoretical justification of the use of constraints. Nevertheless, Michaud's argument is plausible. It would be desirable if one could mitigate the effects of sampling error by some kind of statistical procedure rather than (solely) with bounds.

Michaud (1998) recommends the following: Let μ^H and C^H be the historically observed mean vector and covariance matrix. Assume (for the moment) that these are the true parameters. Also, assume that returns are generated by a joint normal distribution. Based on these assumptions, randomly generate many (say, 500) alternate histories that could have been observed. Each of these 500 alternate "histories" has an "observed" mean return vector and covariance matrix μ^s and C^s, $s = 1, \ldots, 500$. From these, 500 alternate efficient frontiers EF can be computed. One use which Michaud makes of the 500 EFs is to evaluate the means and standard deviations of their portfolios in terms of the original, historic μ^H, C^H and plot all these on one $E\sigma$ graph. This provides a "cloud" of points rather than a single curve. Michaud recommends that if the $E\sigma$ of the investor's current portfolio is within this cloud, e.g., is closer to the efficient frontier than all but the outer 5% of the EFs, then the apparent inefficiency of the current portfolio may be sampling error and the portfolio should not be changed.

Another use which Michaud makes of the 500 EFs is to average the holdings of portfolios which are, in some sense, at corresponding places on each of the 500 frontiers. (See Michaud's specification of the portfolios to be averaged; other specifications could be used without altering the basic idea.) Michaud refers to these averaged portfolios as "*resampled*" efficient portfolios which make up a resampled efficient frontier.[7] Sim-

[7] Resampled efficiency, as described in Michaud (1998), Chapters 6 and 7, was co-invented by Richard Michaud and Robert Michaud and is a U.S. patented procedure, #6,003,018, December 1999, patent pending worldwide. New Frontier Advisors, LLC, has exclusive licensing rights worldwide.

ilarly, the set of $E\sigma$ combinations in the preceding paragraph (whose inefficiency is considered to be quite possibly due to sampling error) is referred to as the resampled confidence region. Michaud presents simulation results that indicate that the resampled efficient portfolios will be more efficient in fact, in terms of a hypothetical true μ^T, C^T than would be an efficient frontier based on one observed μ^H, C^H drawn randomly from μ^T, C^T.

One glaring weakness in Michaud's otherwise persuasive arguments is that he makes much of the fact that the portfolio at the top of the resampled efficient frontier is diversified whereas the top of frontier for any given μ^H, C^H is not diversified. The problem is that the top point on the traditional frontier maximizes E ignoring σ. The fact that the highest point on the frontier is undiversified does not require a new methodology to solve; it only requires picking a portfolio lower on the frontier. It was the undiversified nature of E-maximizing portfolios that lead to the introduction of σ in the first place.

Except for this issue, Michaud raises serious questions about how to treat parameter uncertainty. Markowitz and Usmen (2003) report an experiment comparing Michaud's approach with the one based on Bayesian inference with a "diffuse prior". The authors were surprised to find that their results generally favor the Michaud approach.

10. Epilogue

We have mostly focused on mean-variance in this chapter on risk-return analysis. References have been supplied to more extensive accounts of semivariance, Value at Risk, Conditional Value at Risk, probability of loss and mean absolute deviation measures.

If one subscribes to the Von Neumann–Morgenstern–Savage view of rational action under risk and uncertainty, then mean-variance analysis is to be judged as a convenient, practical approximation to maximizing expected utility. Comparisons between expected utility and mean-variance approximations suggest that the latter are fairly robust.

One source of convenience for mean-variance as compared to utility analysis is that the former does not require the portfolio analyst to determine explicitly the investor's utility function. Another convenience is that only a vector μ of means and a covariance matrix C are required as estimates of the joint return distribution. When measures of risk other than variance (or standard deviation) are used, then historical portfolio risk, e.g., historical mean absolute deviation is typically used as the estimator of forthcoming portfolio risk. This can be done with mean-variance analysis as well by using a historical covariance matrix for C. But when variance is used as the risk criteria one can (more easily than with other measures of risk) use factor, scenario, factor analysis, or Bayesian models for estimating C. We have briefly reviewed some of these, as well as alternate methods of estimating μ.

Another source of convenience for mean-variance as compared to expected utility maximization is speed of computation. This can still be important if an investor is waiting, in real time, for the results of a simulation analysis requiring the computation of many frontiers. When entire efficient frontiers are to be traced out, computation takes

advantage of some neat properties of these frontiers; in particular, in portfolio space the efficient frontier is piecewise linear; in mean-variance space, piecewise parabolic; in mean-standard deviation space the pieces are segments of hyperbola or straight lines. We have presented the formulas for these pieces and have discussed how each efficient segment is connected to its adjacent efficient segments.

These formulas and geometric characterization are true not only for the simple "standard" portfolio analysis model, but also for any instance of the "general" portfolio selection model. The latter permits any system of linear equalities or inequalities as constraints. We have illustrated how such can be used to model various asset/liability situations.

References

Aitken, A.C., 1956. Determinants and Matrices, ninth ed. Oliver and Boyd, Edinburgh and London; Interscience, New York.

Arrow, K., 1971. Aspects of the Theory of Risk Bearing. Markham Publishing Company, Chicago, IL.

Barber, B., Lehavy, R., McNichols, M., Trueman, B., 2001. Can investors profit from the prophets? Security analyst recommendations and stock returns. Journal of Finance 56 (2), 531–563.

Bellman, R.E., 1957. Dynamic Programming. Princeton University Press, Princeton, NJ.

Black, F., 1972. Capital market equilibrium with restricted borrowing. Journal of Business, 444–455, July.

Black, F., Litterman, R., 1991. Asset allocation: Combining investor views with market equilibrium. Journal of Fixed Income 1 (2), 7–18.

Black, F., Litterman, R., 1992. Global portfolio optimization. Financial Analysts Journal 48 (5), 28–43.

Blin, J., Douglas, G., 1987. Stock returns vs factors. Investment Management Review, 36–46, September.

Bloch, M., Guerard, J., Markowitz, H., Todd, P., Xu, G., 1993. A comparison of some aspects of the U.S. and Japanese equity markets. Japan and the World Economy 5, 3–26.

Blume, M.E., 1975. Betas and their regression tendencies. Journal of Finance 30 (3), 785–795.

Blume, M.E., 1979. Betas and their regression tendencies: Some further evidence. Journal of Finance 34 (1), 265–267.

Brinson, G.P., Hood, L.R., Beebower, G.L., 1986. Determinants of portfolio performance. Financial Analysts Journal 42 (4), 39–44.

Brinson, G.P., Hood, L.R., Beebower, G.L., 1995. Determinants of portfolio performance. Financial Analysts Journal 51 (1), 133–138.

Brinson, G.P., Singer, B.D., Beebower, G.L., 1991. Determinants of portfolio performance II: An update. Financial Analysts Journal 47 (3), 40–48.

Chamberlain, G., 1983. A characterization of the distributions that imply mean variance utility functions. Journal of Economic Theory 29, 185–201, June.

Chan, K., Narasimhan, C., Lakonishok, J., 1996. Momentum strategies. Journal of Finance 51 (5), 1681–1713.

Chopra, V.K., Ziemba, W.T., 1993. The effect of errors in means, variances and covariances on optimal portfolio choice. Journal of Portfolio Management 19 (2), 6–12.

Christofi, C., Theodossiou, P., 1995. Simple criteria for optimal portfolio selection revisited. Advances in Mathematical Programming and Financial Planning 4 (1), 45–59.

Cohen, K.J., Pogue, J.A., 1967. An empirical evaluation of alternative portfolio-selection models. Journal of Business 40 (2), 166–193, April.

Dantzig G.B., 1954. Notes on linear programming: Parts VIII IXX—upper bounds, secondary constraints, and block triangularity in linear programming. The RAND Corporation, Research Memorandum RM-1367, October 4, 1954. Published in Econometrica 23 (2) (April 1955) 174–183.

Dantzig, G.B., 1955. Linear programming under uncertainty. Management Science 1, 197–206.

Dantzig, G.B., 1963. Linear Programming and Extensions. Princeton University Press, Princeton, NJ.
Dexter, A.S., Yu, J.N.W., Ziemba, W.T., 1980. Portfolio selection in a lognormal market when the investor has a power utility function: Computational results. In: Dempster, M.A.H. (Ed.), Stochastic Programming. Academic Press, New York, pp. 507–523.
Dimson, E., Marsh, P., Staunton, M., 2002. Triumph of the Optimists. Princeton University Press, Princeton, NJ.
Ederington, L.H., 1986. Mean-variance as an approximation to expected utility maximization. Working Paper 86-5, School of Business Administration, Washington University, St Louis, MO.
Elton, E.J., Gruber, M.J., 1973. Estimating the dependence structure of share prices-implications for portfolio selection. The Journal of Finance XXVII (4), 1203–1233, December.
Elton, E.J., Gruber, M.J., Padberg, M.W., 1976. Simple criteria for optimal portfolio selection. The Journal of Finance 31 (5), 1341–1357, December.
Elton, E.J., Gruber, M.J., Padberg, M.W., 1977. Simple rules for optimal portfolio selection: The multi group case. Journal of Financial and Quantitative Analysis 12 (3), 329–345.
Elton, E.J., Gruber, M.J., Padberg, M.W., 1978. Simple criteria for optimal portfolio selection: Tracing out the efficient frontier. The Journal of Finance 33 (1), 296–302, March.
Fama, E.F., French, K.R., 1995. Size and book-to-market factors in earnings and returns. Journal of Finance 50 (1), 131–155.
Farrell, J.L. Jr., 1983. Guide to Portfolio Management. McGraw-Hill, New York.
Farrell, J.L. Jr., 1985. The dividend discount model: A primer. Financial Analysts Journal 41 (6), 16–19, 22–25, November/December.
Feeney, G.J., Hester, D.D., 1967. Stock market indices: A principal components analysis. In: Hester, D.D., Tobin, J. (Eds.), Risk Aversion and Portfolio Choice. In: Cowles Foundation Monograph, vol. 19. John Wiley & Sons, New York, pp. 110–138.
Feller, W., 1970. An Introduction to Probability Theory and Its Applications, revised printing of third ed. John Wiley and Sons, New York.
Fletcher, R., 1987. Practical Methods of Optimization, second ed. John Wiley and Sons, New York.
Frost, P., Savarino, J., 1988. For better performance: Constrain portfolio weights. Journal of Portfolio Management 15 (1), 29–34.
Gibson, R.C., 2000. Asset Allocation: Balancing Financial Risks, third ed. McGraw-Hill, New York.
Gordon, M., 1962. The Investment, Financing, and Valuation of the Corporation. Irwin, Homewood, IL.
Grauer, R.R., 1986. Normality, solvency, and portfolio choice. Journal of Financial and Quantitative Analysis 21 (3), 265–278, September.
Hakansson, N.H., 1971. On optimal myopic portfolio policies, with and without serial correlation of yields. Journal of Business 44 (3), 324–334.
Haugen, R., Baker, N., 1996. Commonality in the determinants of expected stock returns. Journal of Financial Economics 41 (3), 401–439.
Hlawitschka, W., 1994. The empirical nature of Taylor-series approximations to expected utility. The American Economic Review 84 (3), 713–719.
Infanger, G., 1994. Planning Under Uncertainty Solving Large-Scale Stochastic Linear Programs. Boyd & Fraser Publishing Company, Massachusetts.
Jacobs, B.I., Levy, K.N., 1988. Disentangling equity return regularities: New insights and investment opportunities. Financial Analysts Journal 44 (3), 18–44, May/June.
Jacobs, B.I., Levy, K.N., Markowitz, H.M., 2004. Financial market simulation. Journal of Portfolio Management, 142–152 (30th Anniversary).
Jacobs, B.I., Levy, K.N., Markowitz, H.M., 2005. Portfolio optimization with factors, scenarios, and realistic short positions. Operations Research 53 (4), 586–599, July/August.
Jacobs, B.I., Levy, K.N., Markowitz, H.M., 2006. Trimability and fast optimization of long-short portfolios. Financial Analysts Journal, 36–46, March/April.
Jacobs, B.I., Levy, K.N., Markowitz, H.M., Starer, D., 2006. Equivalence and relative computational complexity of the critical line and Elton–Gruber–Padberg algorithms, Jacobs Levy Equity Management. Working Paper, February.

Jobson, J.D., Korkie, B., 1981. Putting Markowitz theory to work. Journal of Portfolio Management 7 (4), 70–74.

Jorion, P., 1997. Value at Risk. R.D. Irwin, Chicago, IL.

Journal of Investing, 1994. 3 (3), Fall.

Kallberg, J.G., Ziemba, W.T., 1981. Remarks on optimal portfolio selection. In: Bamberg, G., Optiz, O. (Eds.), Methods of Operations Research. Oelgeschlager, Gunn and Hain, Cambridge, MA, pp. 507–520.

Kallberg, J.G., Ziemba, W.T., 1983. Comparison of alternative utility functions in portfolio selection problems. Management Science 29 (11), 1257–1276, November.

Kallberg, J.G., Ziemba, W.T., 1984. Mis-specifications in portfolio selection problems. In: Bamberg, G., Spremann, K. (Eds.), Risk and Capital. Springer-Verlag, New York, pp. 74–87.

King, B.F., 1966. Market and industry factors in stock price behavior, part II. Journal of Business 39 (1), 139–190.

Konno, H., Yamazaki, H., 1991. Mean-absolute deviation portfolio optimization model and its applications to Tokyo stock market. Management Science 37 (5), May.

Koopmans, T.C., 1951. Analysis of production as an efficient combination of activities. In: Koopmans, T.C. (Ed.), Activity of Production and Allocation, seventh ed. Yale University Press, New Haven and London.

Kroll, Y., Levy, H., Markowitz, H.M., 1984. Mean variance versus direct utility maximization. Journal of Finance 39 (1), March.

Kuhn, H.W., Tucker, A.W., 1951. Nonlinear programming. In: Neyman, J. (Ed.), Proceedings of the Second Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, CA, pp. 481–492.

Lakonishok, J., Shleifer, A., Vishny, R.W., 1994. Contrarian investment, extrapolation and risk. Journal of Finance 49 (5), 1541–1578.

Lamont, O., 1998. Earnings and expected returns. Journal of Finance 53, 1563–1588.

Levy, H., Markowitz, H.M., 1979. Approximating expected utility by a function of mean and variance. American Economic Review 69 (3), 308–317, June.

Lintner, J., 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. Review of Economics and Statistics, 13–37, February.

MacKay, R., Keefer, T.E., 1966. VaR is a dangerous technique, Corporate Finance, Searching for Systems Integration Supplement, September.

Markowitz, H.M., 1952. Portfolio selection. The Journal of Finance 7 (1), 77–91, March.

Markowitz, H.M., 1956. The optimization of a quadratic function subject to linear constraints. Naval Research Logistics Quarterly 3, 111–133.

Markowitz, H.M., 1959. Portfolio Selection: Efficient Diversification of Investments. Wiley, Yale University Press, 1970, second ed., Basil Blackwell, 1991.

Markowitz, H.M., 1987. Mean-Variance Analysis in Portfolio Choice and Capital Markets. Basil Blackwell, Cambridge, MA.

Markowitz, H.M., 1999. Foreword to Elton, E.J., Gruber, M.J., Investments: Portfolio Theory and Asset Pricing. MIT Press, Cambridge, MA.

Markowitz, H.M., 2000. Foreword to B.I. Jacobs and K.N. Levy, Equity Management: Quantitative Analysis for Stock Selection. McGraw-Hill, New York.

Markowitz, H.M., 2005. Market efficiency: A theoretical distinction and so what? Financial Analysts Journal 61 (5), 17–30, September/October.

Markowitz, H.M., Perold, A.F., 1981a. Portfolio analysis with factors and scenarios. The Journal of Finance 36 (14), 871–877, September.

Markowitz, H.M., Perold, A.F., 1981b. Sparsity and piecewise linearity in large portfolio optimization problems. In: Duff, I.S. (Ed.), Sparse Matrices and Their Uses. Academic Press, San Diego, CA, pp. 89–108.

Markowitz, H.M., Todd, P., 2000. Mean-Variance Analysis in Portfolio Choice and Capital Markets. F. Fabozzi and Associates, New Hope, PA. Reissue of Markowitz (1987) with chapter by P. Todd.

Markowitz, H.M., Usmen, N., 2003. Resampled frontiers versus diffuse bayes: An experiment. Journal of Investment Management 1 (4), 9–25.

Markowitz, H.M., van Dijk, E., 2003. Single-period mean-variance analysis in a changing world. Financial Analysts Journal 59 (2), 30–44, March/April.
Markowitz, H.M., Reid, D.W., Tew, B.V., 1994. The value of a blank check. The Journal of Portfolio Management, 82–91, Summer.
Markowitz, H.M., Todd, P., Yamane, Y., Xu, G., 1993a. Fast computation of mean-variance efficient sets using historical covariances. Journal of Financial Engineering 1 (2), 117–132.
Markowitz, H.M., Todd, P., Yamane, Y., Xu, G., 1993b. Computation of mean-semivariance efficient sets by the critical line algorithm. Annals of Operations Research 45, 307–317.
Mausser, H., Rosen, D., 1999. Beyond VaR: From measuring risk to managing risk. ALGO Research Quarterly 1 (2), 5–20.
Merton, R.C., 1973. An intertemporal capital asset pricing model. Econometrica 41, 867–887, September.
Michaud, R.O., 1989. The Markowitz optimization enigma: Is optimized optimal? Financial Analysts Journal 45 (1), 31–42.
Michaud, R.O., 1998. Efficient Asset Management: A Practical Guide to Stock Portfolio Optimization and Asset Allocation. Harvard Business School Press, Boston, MA.
Mossin, J., 1966. Equilibrium in a capital asset market. Econometrica 34 (4), 768–783, October.
Mossin, J., 1968. Optimal multiperiod portfolio policies. Journal of Business 1 (2), 215–229.
Ohlson, J.A., 1975. The asymptotic validity of quadratic utility as the trading interval approaches zero. In: Ziemba, W.T., Vickson, R.G. (Eds.), Stochastic Optimization Models in Finance. Academic Press, New York, pp. 221–234.
Ohlson, J.A., 1979. Risk, return, security-valuation and the stochastic behavior of accounting numbers. Journal of Financial and Quantitative Analysis 14 (2), 317–336.
Ortobelli, S., Huber, I., Schwartz, E., 2002. Portfolio selection with stable distributed returns. Mathematical Methods of Operations Research 55, 265–300.
Owen, J., Rabinovitch, R., 1983. On the class of elliptical distributions and their applications to the theory of portfolio choice. Journal of Finance 38 (3), 745–752, June.
Perold, A.F., 1984. Large-scale portfolio optimization. Management Science 30 (10), 1143–1160, October.
Penman, S.H., 1996. The articulation of price-earnings ratios and market-to-book ratios and the evaluation of growth. Journal of Accounting Research 34, 235–259.
Pratt, J.W., 1964. Risk aversion in the small and in the large. Econometrica 32, 122–136, January.
Pulley, L.M., 1981. A general mean-variance approximation to expected utility for short holding periods. Journal of Financial and Quantitative Analysis 16, 361–373.
Pulley, L.M., 1983. Mean-variance approximations to expected logarithmic utility. Operations Research 31 (4), 685–696.
Rockafellar, R.T., Uryasev, S., 2000. Optimization of conditional value-at-risk. Journal of Risk 2 (3), 21–41.
Rosenberg, B., 1974. Extra-market components of covariance in security returns. Journal of Financial and Quantitative Analysis 9 (2), 263–273.
Rosenberg, B., Reid, K., Lanstein, R., 1984–85. Persuasive evidence of market inefficiency. Journal of Portfolio Management 11 (3), 9–17.
Ross, S., 1976. The arbitrage theory of capital asset pricing. Journal of Economic Theory 13 (3), 341–360.
Roy, A.D., 1952. Safety first and the holding of assets. Econometrica 20, 431–449.
Samuelson, P.A., 1969. Lifetime portfolio selection by dynamic stochastic programming. The Review of Economics and Statistics 51, 239–246.
Samuelson, P.A., 1970. The fundamental approximation theorem of portfolio analysis in terms of means, variances and higher moments. The Review of Economic Studies 37, 537–542.
Savage, L.J., 1954. The Foundations of Statistics, second revised ed. John Wiley & Sons, Dover, New York.
Schwartz, S.L., Ziemba, W.T., 2000. Predicting returns on the Tokyo stock exchange. In: Keim, D.B., Ziemba, W.T. (Eds.), Security Market Imperfections in Worldwide Equity Markets, Cambridge, MA, pp. 492–511.
Sharpe, W.F., 1963. A simplified model for portfolio analysis. Management Science 9 (2), 277–293, January.
Sharpe, W.F., 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. The Journal of Finance 19 (3), 425–442, September.

Sharpe, W.F., Tint, L.G., 1990. Liabilities—a new approach. Journal of Portfolio Management 16 (2), 5–10, Winter.

Simaan, Y., 1993. What is the opportunity cost of mean-variance investment strategies? Management Science 39 (5), 578–587, May.

Sortino, F., Satchell, S., 2001. Managing Downside Risk in Financial Markets: Theory, Practice and Implementation. Butterworth–Heinemann, Burlington, MA.

Swensen, D.F., 2000. Pioneering Portfolio Management: An Unconventional Approach to Institutional Investment. Free Press, New York.

Tobin, J., 1958. Liquidity preference as behavior towards risk. Review of Economic Studies 25 (1), 65–86, February.

Von Neumann, J., Morgenstern, O., 1944. Theory of Games and Economic Behavior. Princeton University Press, Princeton, NJ. Third ed. in 1953.

Williams, J.B., 1938. The Theory of Investment Value. Harvard University Press, Cambridge, MA.

Worzel, K.J., Vassiadou-Zeniou, C.C., Zenios, S.A., 1994. Integrated simulation and optimization models for tracking indices of fixed-income securities. Operations Research 42 (2), 223, March–April.

Young, M.R., 1998. A minimax portfolio selection rule with linear programming solution. Management Science 44 (5), 673, May.

Young, W.E., Trent, R.H., 1969. Geometric mean approximation of individual security and portfolio performance. Journal of Financial Quantitative Analysis 4, 179–199, June.

Ziemba, W.T., 1974. Choosing investment portfolios when the returns have stable distributions. In: Hammer, P.L., Zoutendijk, G. (Eds.), Mathematical Programming in Theory and Practice. North-Holland Publ., Amsterdam, pp. 443–482.

Ziemba, W.T., Mulvey, J.M. (Eds.), 1998. Worldwide Asset and Liability Modeling. Cambridge University Press, Cambridge, MA.

Zenios, S.A., Kang, P., 1993. Mean absolute deviation optimization for mortgage-backed securities. Annals of Operations Research 45, 433–450.

This page intentionally left blank

Chapter 5

DYNAMIC ASSET ALLOCATION STRATEGIES USING A STOCHASTIC DYNAMIC PROGRAMMING APPROACH

GERD INFANGER

Department of Management Science and Engineering, Stanford University, Stanford, CA 94305-4026, USA
and
Infanger Investment Technology, LLC, 2680 Bayshore Parkway, Suite 206, Mountain View, CA 94043, USA

Contents

Abstract 200
Keywords 200
1. Introduction 201
2. Approaches for dynamic asset allocation 204
   2.1. Multi-stage stochastic programming 204
   2.2. Stochastic dynamic programming 206
3. Single-period portfolio choice 207
4. Utility functions 209
5. A general approach to modeling utility 211
6. Dynamic portfolio choice 214
   6.1. Dynamic stochastic programming and Monte Carlo sampling 215
   6.2. Serially dependent asset returns 216
   6.3. A fast method for normally distributed asset returns 217
7. Numerical results 217
   7.1. Data assumptions 217
   7.2. An investment example 222
   7.3. The performance of dynamic strategies 234
   7.4. Dynamic strategies for hedging downside risk 238
   7.5. Downside risk protection at every period 241
   7.6. Computation times 246
8. Comparison to multi-stage stochastic programming 247
Acknowledgements 248
References 248

Handbook of Asset and Liability Management, Volume 1
Edited by S.A. Zenios and W.T. Ziemba
© 2006 Published by Elsevier B.V.
DOI: 10.1016/S1872-0978(06)01005-2

Abstract

A major investment decision for individual and institutional investors alike is to choose between different asset classes, i.e., equity investments and interest-bearing investments. The asset allocation decision determines the ultimate risk and return of a portfolio. The asset allocation problem is frequently addressed either through a static analysis, based on Markowitz' mean-variance model, or dynamically but often myopically through the application of analytical results for special classes of utility functions, e.g., Samuelson's fixed-mix result for constant relative risk aversion. Only recently, the full dynamic and multi-dimensional nature of the asset allocation problem could be captured through applications of stochastic dynamic programming and stochastic programming techniques, the latter being discussed in various chapters of this book. The chapter reviews the different approaches to asset allocation and presents a novel approach based on stochastic dynamic programming and Monte Carlo sampling that permits one to consider many rebalancing periods, many asset classes, dynamic cash flows, and a general representation of investor risk preference. The chapter further presents a novel approach of representing utility by directly modeling risk aversion as a function of wealth, and thus provides a general framework for representing investor preference. The chapter shows how the optimal asset allocation depends on the investment horizon, wealth, and the investor's risk preference and how it therefore changes over time depending on cash flow and the returns achieved. The chapter demonstrates how dynamic asset allocation leads to superior results compared to static or myopic techniques. Examples of dynamic strategies for various typical risk preferences and multiple asset classes are presented.

Keywords

portfolio theory and applications, dynamic asset allocation, stochastic dynamic programming, stochastic programming

JEL classification: C61, D81, G1

1. Introduction

The major decision of an investor regarding his/her portfolio is to choose the allocation between different asset classes, especially between equity investments and interest-bearing investments. Strategic asset allocation determines the ultimate expected rate of return and risk of an investor's portfolio, see, e.g., Blake, Lehmann and Timmermann (1999). For individuals and institutional investors alike, it is usually a long-term decision and its analysis should include all financial aspects, e.g., wealth, current and future cash flows, and financial goals. Inflation, as well as liquidity considerations (to plan for the unexpected), should be part of the analysis.

In the long run, equity investments have grown at a faster rate than other assets such as bonds, T-bills, and inflation, see, e.g., Siegel (2002), Constantinides (2002), and Dimson, Marsh and Staunton (2002). However, in the short run, the risk is significant and may lead to negative returns. Even in the long run, equity investments can be quite risky, and as we have painfully observed during the recent downturn from 2000 till 2003 as well as in bear markets before, equity investments can quickly lose a significant portion of their value. interest-bearing investments have exhibited lower returns in the long run, barely exceeding inflation, but with less risk, making them possibly a better vehicle for the short-term. It is clear that one needs to determine the right balance between the asset classes, and this balance depends on how much risk an investor is willing to assume and may change over time due to changes in wealth, needs, goals and market conditions. It is well known that strategic asset allocation has a far greater impact on the overall performance of an investor's portfolio than, for example, the selection of individual securities.

The question of how the length of the investment horizon impacts the optimal asset allocation is an important theoretical and practical question, which has been discussed in the academic literature for more than 30 years. The answers to this question vary significantly depending on the assumptions made. For example, Levy (1972) and Gunthorpe and Levy (1994) discussed portfolio performance and composition versus investment horizon in a mean-variance framework. Practitioners tend to recommend a larger allocation towards stocks as the investment horizon increases. This is often argued under the name of "time diversification", and is based on the argument that if stocks are distributed independently and identically in each time period according to a log normal distribution, the distribution over many periods as the product of log normals is also log normal and the mean and the variance of the logarithm of the return distribution grows proportionally with the length of the investment period. Since the standard deviation of the log returns then grows with the square root of the horizon, the probability of capital loss decreases as time increases, and stocks become more favorable as the time horizon increases. Using this argument with a reasonable assumption of a 10% mean and 15% volatility, annually, one may verify that after an investment horizon of about seven years, the return on stocks exceeded with 95% probability is positive. It would take 13 years to arrive at a positive return exceeded with 99% probability.

Time diversification often includes a broader view of human capital, suggesting that the capital represented by an individual's ability to work and generate wages should also enter the equation. Consequently, at the beginning of one's professional life, most of one's capital is the potential of generating future income through labor, which is considered to be not risky, while at the end of one's professional career, most capital lies in one's financial assets (in the retirement account, but also real estate, stocks, etc.), and the ability to work and to generate income becomes a lesser part of the total capital. Thus, it is argued that at the beginning of one's career, the small amount in financial assets should be invested in a more risky allocation and gradually reduced to a less risky allocation at retirement, when all assets are the financial assets generated during the lifetime. Based on time diversification arguments, a common rule of thumb used by practitioners is to invest in stocks the percentage resulting from 100-age, thus a 30 year old should invest about 70% of his portfolio into stocks, and a 60 year old about 40%. These arguments only use time as the explaining factor and the corresponding strategies do not react to investment performance.

Treynor (2003) on time diversification puts forward the question of how each year's investment choices influence the wealth accumulated at the end of one's career. This perspective concludes that the impact of each year's dollar gains and losses on terminal wealth depend on the riskless rate, and one should time diversify in such a way that the risk evaluated in terminal dollars should be constant over the investment horizon. Therefore, unless investors can predict different returns for each year, the money amount exposed to the stock market should be approximately constant over the life-time.

Samuelson (1969) in his landmark work addressed how much of his/her wealth an individual should either consume or invest at any period in time, assuming no bequest is to be left behind. He proves using a backward dynamic programming recursion that, given the choice of one risky asset and one risk free asset, for the returns of the risky asset distributed identically and independently (iid) over time, for all income generated through investment, and for individuals valuing their consumption in time according to a power utility function with respect to consumption and maximizing discounted expected utility over the lifetime, it is optimal to invest the same proportion of wealth into stocks in every period, independently of wealth. The same was proved by Merton (1969) in continuous time and later, see Merton (1990), extended to multiple risky assets and various bequest situations. This life-time portfolio and consumption selection prompted an apparent conflict between theoreticians and practitioners, since the advice from Samuelson and Merton is quite different from what financial practitioners tell their clients. The remarkable aspect of Samuelson's and Merton's result is that, under their assumptions about the market and under constant relative risk aversion, the consumption decisions and the investment decisions are independent of each other, and therefore the optimal investment decision is not only invariant with respect to investment horizon but also with respect to wealth. Thus, the result translates directly to the investment problem only, where one wants to maximize the utility of final wealth at the end of the investment horizon, by allocating and re-allocating at each period along the way. The

result follows directly from the utility function used, stipulating that the (relative) risk aversion of the individual is invariant with respect to wealth.

Optimal multi-period investment strategies based on maximizing expected utility have been analyzed by Mossin (1968). Mossin's research attempted to isolate the class of utility functions of terminal wealth which result in myopic utility (value) functions for intermediate wealth positions. For a discussion of utility functions, see Section 4 below. Myopic means that these functions are independent of returns beyond the current period. Thus, it would be sufficient to analyze only the current optimization problem to arrive at an optimal multi-period investment strategy. Mossin concluded that logarithmic functions (for general asset distributions) and power functions (for serially independent asset distributions) are completely myopic. He also contended that, if there is a riskless asset (whose return is known for the entire investment horizon), and a risky one, all utility functions for which the risk tolerance is linear in wealth (HARA) would lead to partial myopia as one would optimally invest in any period as if in all further periods the investment were to be made in the riskless asset, and complete myopia would exist if the risk-free rate were to be zero. However, Hakansson (1971) demonstrated that, for the HARA case, even when asset returns are serially independent, no myopic strategies are optimal, except for the highly restricted case of a complete absence of any restrictions on short sales and borrowing. A percentage margin requirement, an absolute limit on borrowing, or a reasonable lending constraint such that the borrowed money would have to be repaid would not lead to a myopic strategy. Thus, in the presence of such restrictions only the power and logarithmic utility functions would lead to myopia. Furthermore, if asset returns are serially correlated, only the logarithmic utility function would result in a myopic policy. Later, Cox and Huang (1999) presented an analytical solution based on diffusion processes for the consumption-investment problem with HARA utility function and wealth and consumption constrained to be nonnegative. Later, an approach using approximate analytical solutions has been developed by Campbell and Viciera Campbell and Viceira (2002), based on perturbations of known exact solutions. See also the chapter by Chacko and Neumar (2006) in this book.

Thus we can summarize, a logarithmic utility function results in a myopic portfolio strategy, both for serially dependent and independent asset return distributions; a power utility function results in a myopic strategy only in the case of serially independent asset returns distributions; and a HARA utility function results in a myopic strategy only for serially independent asset return distributions and only in a non-realistic setting of complete absence of borrowing and short-selling constraints. All other utility functions do not result in myopic investment strategies for any return distributions.

More recently, numerical dynamic portfolio optimization methods have been developed that permit one to determine the asset allocation strategy that maximizes an investor's expected utility. These new approaches are based on stochastic dynamic programming and stochastic programming and promise to accurately solve for various types of utility functions and asset return processes.

2. Approaches for dynamic asset allocation

The two major approaches successful in solving practical dynamic asset allocation problems are stochastic dynamic programming (stochastic control) and stochastic programming. As we have discussed above, asset allocation problems under restrictive assumptions can be solved analytically. In practice, fixed-mix strategies are commonly implemented and often lead to very good results.

2.1. Multi-stage stochastic programming

The stochastic programming approach can efficiently solve the most general models, where transaction costs need to be considered, and the returns distributions have general serial dependency. The stochastic programming approach also lends itself well to the more general asset liability management problem (ALM). Here, liabilities in addition to assets need to be considered. This problem is faced by pension funds and insurance companies. Besides assets, pension plans need to consider retirement obligations, which may depend on uncertain economic and institutional variables, and insurance companies need to consider uncertain pay-out obligations due to unforeseen and often catastrophic events. Asset liability models are most useful when both asset returns and liability pay-outs are driven by common, e.g., economic, factors. In this case ALM represents the only approach that can take into account directly the joint distribution of asset returns and liability cash flows. Lenders operating in the secondary mortgage market also face a certain kind of ALM problem, when deciding on re-financing their (pools of) mortgages by issuing a portfolio of bonds, callable or non-callable with various maturities. Here the assets are the mortgages bought from banks and the liabilities are the bonds issued, see, e.g., Infanger (1999).

Traditional stochastic programming uses scenario trees to represent possible future events. The trees may be constructed by a variety of scenario generation techniques, with the emphasis on keeping the resulting tree thin but representative of the event distribution in order to arrive at a computationally tractable problem. Often, in later decision stages of the model, only a very small number of scenarios is used as a representation of the distribution leading to very thin sub-trees. Thus, the emphasis is on obtaining a good first-stage solution rather than obtaining an entire accurate policy. Early applications of stochastic programming for asset allocation are discussed in Mulvey and Vladimirou (1992), formulating financial networks, Golub et al. (1995). Early applications of stochastic programming for dynamic fixed-income strategies are discussed in Zenios (1993), discussing the management of mortgage-backed securities, Hiller and Eckstein (1993), and Nielsen and Zenios (1996). Early practical applications of stochastic programming for asset liability management are reported in Kusy and Ziemba (1986) for a bank and in Carino et al. (1994) for an insurance company. Ziemba (2003) gives a summary of the stochastic programming approach for asset liability management. An approach based on partitioning the probability space and calculating deterministic

bounds was developed by Frauendorfer (1996), and used for bond management and asset liability management. The book edited by Wallace and Ziemba (2005) gives publicly available stochastic programming code. Stochastic programming software can be best used from within a modeling system, for example, GAMS (Brooke, Kendrik and Meeraus, 1988) provides DECIS (Infanger, 1997) as an integrated stochastic programming solver.

Monte Carlo sampling is an efficient way to go when representing multi-dimensional distributions. An approach, referred to as decomposition and Monte Carlo sampling, uses Monte Carlo (importance) sampling within a decomposition for estimating Benders cut coefficients and right-hand sides. This approach has been developed by Dantzig and Glynn (1990) and Infanger (1992). Dantzig and Infanger (1993) show how the approach of decomposition and importance sampling could be used for solving multi-period asset allocation problems. The success of the sampling within the decomposition approach depends on the type of serial dependency of the stochastic parameter processes, determining whether or not cuts can be shared or adjusted between different scenarios of a stage. Infanger (1994) and Infanger and Morton (1996) show that for serial correlation (in form of autoregressive processes) of stochastic parameters, unless in the right hand side of the (linear) program, cut sharing is difficult for more than three decision stage problems. However, for serially independent stochastic parameters, the approach is very efficient for solving problems with many decision stages.

Monte Carlo pre-sampling uses Monte Carlo sampling to generate a tree, much like the scenario generation methods referred to above, and then employs a suitable method for solving the sampled (and thus approximate) problem. Infanger (1999) used the pre-sampling approach for representing the mortgage funding problem. This approach combines optimization and simulation techniques to represent a 360 month problem by four decision stages (initial, after a year, after 5 years, and at the end of the horizon) that are subject to optimization and by pre-defined decision rules representing intermediate decisions. The paper also provides an efficient way to independently evaluate the solution strategy as a result from solving the multi-stage stochastic program to obtain a valid upper bound on the objective. The pre-sampling approach is general in terms of modeling and solving stochastic processes with serial dependency; however, it is limited in the number of decision stages, since problems with many decision stages become computationally difficult. Assuming a reasonable sample size for representing the decision tree, problems with up to four decision stages are meaningfully tractable. Thus, if one were to represent asset allocation problems with many time (rebalancing) periods, one needed to represent more than one time period in one decision stage and define rules as to how to manage the assets and liabilities between decision stages. For example, one could assume buy and hold between decision stages and allow for rebalancing at the decision stages. In many situations this is considered a sufficiently accurate approximation of all future recourse decisions. A stochastic programming approach using pre-sampling has been employed by Collomb and Infanger (2005) to analyze the impact of serial dependency on the solution of dynamic asset allocation problems. Instead of pre-defining rebalancing periods up-front, the rebalancing decision may be modeled as

depending on certain conditions occurring. For example, MacLean, Zhao and Ziemba (2006) model portfolio rebalancing as conditioned on prices exceeding or going below certain levels, and MacLean, Ziemba and Li (2005) model portfolio rebalancing when certain wealth goals are met.

In this book, addressing dynamic strategies using stochastic programming for large institutions, Kouwenberg and Zenios (2006) introduce stochastic programming models as they apply to financial decision making, Mulvey et al. (2006) and Ziemba (2006) discuss the application of stochastic programming for multinational insurance companies, Consiglio, Cocco and Zenios (2006) discuss the application of stochastic programming for insurance products with guarantees, Edirishinghe (2006) discusses dynamic strategies for money management, and Consigli (2006) discusses the development of stochastic programming models for individual investors.

2.2. Stochastic dynamic programming

When the focus is on obtaining optimal policies and transaction costs are not the primary issue, stochastic dynamic programming proves to be a very effective approach. Stochastic dynamic programming based on Bellman's (1957) dynamic programming principle has been used for the derivation of the theoretical results obtained for the HARA class of utility functions discussed above. For general monotone increasing and concave utility functions, no analytical solutions are available. However, stochastic dynamic programming can be used as an efficient numerical algorithm when the state space is small, say, up to three or four state variables. This limitation in the number of state variables is well known as the "curse of dimensionality" in dynamic programming. Recently, new methods of value function approximations, see, e.g., De Farias and Van Roy (2003) show promise for problems with larger state spaces; however, it is unclear at this point in time how accurately these methods will approximate the solution of the asset allocation problem. Stochastic dynamic programming as a numerical algorithm has been used by Musumeci and Musumeci (1999), representing results with two asset classes, one stock index and a risk free asset, where in the dynamic programming procedure they condition on the amount of wealth invested in the risky asset. Earlier, Adaci (1996) (in a Ph.D. thesis supervised by George Dantzig and the author) conditioned on wealth and thus set the stage for multiple asset classes, but reported results also only for two asset classes and a few periods. Brennan, Schwartz and Lagnado (1997, 1998) proposed a dynamic model using discrete state approximations including four state variables.

In this chapter we develop an efficient approach for solving the asset allocation problem under general utility functions, many time periods (decision stages), and many asset classes. We next review single-period portfolio choice and draw the connections between Markovitz' mean-variance analysis and utility maximization. We then discuss the properties of various utility functions and present a general framework for modeling utility. Then we discuss multi-period portfolio choice and present a novel approach based on stochastic dynamic programming and Monte Carlo sampling.

3. Single-period portfolio choice

Markowitz (1952), in his seminal work, see also Markowitz and van Dijk (2006) in this book, pointed out that the returns of a portfolio are random parameters and that for the evaluation of a portfolio, one should consider both its expected returns and its risk, where for representing risk he used the portfolio's variance. His mean-variance analysis laid the foundations for modern finance and our understanding as to how markets work. Mean-variance analysis is referred to as modern portfolio theory, whereas post-modern portfolio theory considers extensions including skewed distributions and asymmetric risk measures.

Let $x = (x_1, \ldots, x_n)$ be the holdings of n asset classes under consideration for a portfolio in a certain time period, and let $R = (R_1, \ldots, R_n)$ denote the random rates of return of the asset classes, with mean returns $r = (r_1, \ldots, r_n)$, and covariance matrix $\Sigma = [\sigma_{i,j}]$, where $\sigma_{i,j} = E(R_i - r_i)(R_j - r_j)$ for $i, j = 1, \ldots, n$. Markowitz's mean-variance model is usually stated as

$$\begin{aligned}
&\min \operatorname{var}\left(R^T x\right) = x^T \Sigma x,\\
&e^T x = 1,\\
&r^T x \geqslant \rho,\\
&x \geqslant 0,
\end{aligned}$$

where one chooses the holdings that minimize the variance of the portfolio returns for a given level of expected return ρ. All portfolios selected by the mean-variance model lie on the mean-variance efficient frontier, i.e., increasing the expected return by moving from one efficient portfolio to another means risk (measured through variance) also increases, or smaller risk can only be accomplished by sacrificing expected return.

If a risk-free return is present, and for a single-period analysis this is usually assumed to be the case, then the efficient frontier is represented by the line that intersects the ordinate at the level of the risk-less rate of return and is tangent to the efficient frontier established without the risk-less asset. The portfolio at the point at which the so called market line intersects with the original efficient frontier is the market portfolio. The striking result of Tobin's two fund separation theorem (see Tobin (1958)) is that, if a riskless asset is present, all investors would choose the same portfolio of risky assets, namely the market portfolio, in combination with various amounts of the risk free asset. An investor who wishes to assume the market risk would choose the market portfolio. An investor who is more risk averse than the market would chose a positive fraction of the riskless asset and invest the remainder into the market portfolio. An aggressive investor bearing more risk than the market would borrow at the risk-free rate in addition to his funds at hand and invest the entire amount in the market portfolio, thereby leveraging his wealth. There would be no need for any portfolio other than the market portfolio. For a computation of Tobin's market line and which point on this line to choose, see, Ziemba, Parkan and Brooks-Hill (1974). Usually in a single-period analysis, cash (or money market) represented by the 30-day rate or the three-month rate is considered as

risk-free. There may still be inflation risk, and when considering a single-period horizon of one year, which is a typical investment horizon, the riskless rate may change. However, for a single-period analysis these risks are often considered small enough to be neglected.

Utility theory according to Bernoulli (1738) and Von Neumann and Morgenstern (1944), as a means of dealing with uncertainty, is a generally accepted concept in finance. An investor values different outcomes of a uncertain quantity according to his (Von Neumann–Morgenstern) utility function and maximizes expected utility. The Von Neumann–Morgenstern utility function of a risk-averse investor is a concave function of wealth.

The single-period Markowitz investment problem may also be stated as

$$\begin{aligned} &\max r^T x - \frac{\lambda}{2} x^T \Sigma x, \\ &e^T x = 1, \\ &x \geqslant 0, \end{aligned}$$

trading off expected return and variance in the objective function, where the parameter λ represents the risk aversion of the investor. In this case the investor trades off linearly the expected return and variance of the portfolio, and λ determines how many units of expected return the investor is willing to give up for a decrease of variance by one unit. In terms of maximization of expected utility, the optimization is stated as

$$\begin{aligned} &\max Eu\big(R^T x\big), \\ &e^T x = 1, \\ &x \geqslant 0, \end{aligned}$$

where $u(W)$ is a concave utility function. We maximize the expected utility of wealth at the end of the period, where the initial wealth is normalized to one and the distribution of wealth at the end of the period is $W = R^T x$. Choosing $u(W) = -\exp(-\lambda W)$ (an exponential utility function with risk aversion λ), and assuming the asset returns are distributed as multivariate normal (i.e., $R = N(r, \Sigma)$, with r as the vector of mean returns and Σ the covariance matrix), we can integrate using the exponential transform and obtain $E(-\exp(-\lambda W)) = -\exp(\lambda EW - \frac{\lambda^2}{2}\operatorname{var}(W))$. The certainty equivalent wealth is defined as the fixed amount of wealth that has the same utility as the expected utility of the wealth distribution. Denoting the certainty equivalent wealth as W_c, we can evaluate $-\exp(-\lambda W_c) = -\exp(\lambda EW - \frac{\lambda^2}{2}\operatorname{var}(W))$ and obtain $W_c = EW - \frac{\lambda}{2}\operatorname{var}(W)$. In the case of multivariate normally distributed asset returns and exponential utility, the mean-variance model maximizes certainty equivalent wealth and therefore indirectly expected utility of wealth. For exponential utility and multivariate normally distributed asset returns, maximizing expected utility of wealth and trading off mean versus variance are therefore equivalent.

For normally distributed returns the scope is even broader. For multivariate normally distributed returns and for any monotonically increasing concave utility function $u(W)$, $u'(W) > 0$, $u''(W) < 0$, the obtained optimal solution is mean-variance efficient, and therefore lies on the efficient frontier; see, e.g., Hanoch and Levy (1969). If asset returns are not multi-variate normally distributed, at least one choice of utility function would lead to the equivalent result with respect to mean-variance analysis: $u(W) = W - \frac{\lambda}{2}(W - EW)^2$, namely a quadratic utility function explicitly specifying utility as wealth minus risk aversion times half the square of the deviation around the mean. Summarizing, if the asset return distribution is defined by its first two cumulants only, and all higher cumulants are zero (normal distribution), any monotonically increasing concave utility function will result in a mean-variance efficient solution. If the asset return distribution is characterized by higher (than the first two) non-zero cumulants also, only a quadratic utility function gives a mean-variance efficient solution by explicitly considering only the first and second cumulant of the distribution. Thus, when asset returns are not normally distributed, the optimal solution of an expected utility maximizing problem, except for the quadratic utility case, does not necessarily lie on the mean-variance efficient frontier. Assuming asset returns to be normally distributed, therefore, is very convenient. Also, any linear combination of normally distributed random variables is a normally distributed random variable as well. A theory for choosing portfolios when asset returns have stable distributions is given by Ziemba (1974).

4. Utility functions

A Von Neumann–Morgenstern utility function $u(W)$ represents an investor's attitude towards risk. According to Arrow (1971) and Pratt (1964), the absolute and relative risk aversion coefficients, defined as

$$\mathrm{ARA}(W) = -\frac{u''(W)}{u'(W)}, \qquad \mathrm{RRA}(W) = -W\frac{u''(W)}{u'(W)}$$

determine investor behavior. Thus, it is not the absolute value of the function but how strongly it bends that determines differences in investor choice. A linear utility function representing risk-neutral behavior would exhibit an absolute risk aversion coefficient of zero. Concave utility functions, representing risk-averse behavior, exhibit positive absolute risk aversion. When modeling investor choice, we are interested in how the risk aversion changes with wealth. The inverse of the risk aversion is referred to as the risk tolerance, i.e., $\mathrm{ART}(W) = 1/\mathrm{ARA}(W)$ and $\mathrm{RRT}(W) = 1/\mathrm{RRA}(W)$. Table 1 presents the absolute and relative risk aversion and risk tolerance for some commonly used utility functions.

The power utility function and the logarithmic utility function with $\alpha = 0$ have relative risk aversion that is constant with respect to wealth. They are therefore also referred to as CRRA (constant relative risk aversion) utility functions. The exponential utility function exhibits constant absolute risk aversion with respect to wealth, and is therefore

Table 1
Commonly used utility functions (HARA)

Type	Function	ARA	RRA	ART	RRT
Exponential	$u(W) = -\exp(-\lambda W)$	λ	λW	$\frac{1}{\lambda}$	$\frac{1}{\lambda W}$
Power	$u(W) = \frac{W^{1-\gamma}-1}{1-\gamma}, \gamma > 1$	$\frac{\gamma}{W}$	γ	$\frac{W}{\gamma}$	$\frac{1}{\gamma}$
Generalized Log	$u(W) = \log(\alpha + W)$	$\frac{1}{\alpha+W}$	$\frac{W}{\alpha+W}$	$\alpha + W$	$\frac{\alpha+W}{W}$

said to be of type CARA (constant absolute risk aversion). These utility functions are part of and exhaust the HARA (hyperbolical absolute risk aversion) class, defined by the differential equation $-u'(W)/u''(W) = a + bW$, where a and b are constants; see e.g., Rubinstein (1976).

In a single-period investment problem of one risky asset and one risk-free asset, the amount invested in the risky asset is proportional to the absolute risk tolerance ART(W), and the fraction of wealth invested in the risky asset is proportional to the relative risk tolerance RRT(W). Thus CARA implies that the amount of wealth invested in the risky asset is constant with respect to wealth, and therefore the fraction of the wealth invested in the risky asset declines with increasing wealth proportional to $1/W$. In contrast, CRRA implies that the fraction of wealth invested in the risky asset is constant, and the amount of wealth invested into the risky asset increases proportional to W. Different utility functions exhibit different risk aversion as functions of wealth. Kallberg and Ziemba (1983) discuss the implications of various utility functions in a single-period setting, showing that for concave utility functions and normally distributed asset returns, when the average risk aversion is the same for two utility functions, then the portfolios are also very similar, irrespective of the specific utility functional form. They further argue that under these assumptions any concave increasing utility function could be locally approximated by a quadratic function. In a multi-period investment problem, the relationship between relative and absolute risk aversion and portfolio choice is not as straightforward, and complex investment strategies arise.

Since the risk aversion as a function of wealth reflects investor behavior, the choice of utility function is an important aspect of the modeling. In many situations the HARA class appears too restrictive, and other utility functions have been explored. Bell (1988) defined a class of utility functions satisfying a "one switch" rule. Comparing the preference between two gambles (say, portfolios) at different wealth levels, a switch occurs in that below a certain wealth level the first gamble is preferred and above that wealth level the other. The class of one-switch utility functions when maximizing utility thus would not lead to solutions where for even larger wealth levels, the first gamble would be preferred again, or gambles could switch in and out of preference. His one-switch utility functions include and exhaust the following types: the quadratic ($u(W) = aW^2 + bW + c$), the double exponential (sumex) functions ($U(W) = ae^{bW} + ce^{dW}$), the linear plus exponential ($u(W) = aW + be^{cW}$), and the

linear times exponential ($u(W) = (aW + b)e^{cW}$). Musumeci and Musumeci (1999) found a linear combination of a power and an exponential function attractive, suggesting $u(W) = -W^{-1} - ae^{-bW}$, and give two sets of numerical values for the parameters representing typical more or less risk-averse investors: $a = 2.9556e{-}5$, $b = 8.0e{-}8$ would represent the less, and $a = 1.4778e{-}5$, $b = 4.0e{-}8$ the more risk-averse investor class.

A variant of the quadratic utility function, where investors are concerned with the downside risk of wealth falling below a certain target W_d, is $u(W) = W - \lambda_q/2\,(\max(0, W_d - W))^2$, reflecting a lower partial moment, where risk is not the total variance but the variance of that part conditioned on the wealth falling below the target. We also refer to this utility function as a quadratic downside risk utility function. The linear downside risk utility function, $u(W) = W - \lambda_l \max(0, W_d - W)$, reflects the first lower partial moment, and risk is the expected wealth conditioned on the wealth falling below the target W_d. The efficient frontier with risk as the first lower partial moment is also referred to as the "put–call efficient frontier"; see, e.g., Dembo and Mausser (2000). More generally, the additive function of linear and quadratic downside risk, i.e., $u(W) = W - \lambda_l(\max(0, W_d - W)) - \lambda_q/2(\max(0, W_d - W))^2$, penalizes downside risk, where for $\lambda_l > 0$ the function is non-differentiable at W_d, leading to a jump in risk aversion right at the downside of the target.

It has been generally believed that individual investor utility should be of decreasing absolute risk aversion and increasing relative risk aversion, spanning the range between CARA seen as possibly too conservative for large levels of wealth and CRRA seen as possibly too aggressive for large wealth levels. A classical empirical study about investor preference, Friend and Blume (1975), concludes that for representing average investor preference constant relative risk aversion (CRRA) reflects a first approximation, but there are many deviations in that investors could be exhibiting either increasing or decreasing relative risk aversion. Markowitz (1994) argues, based on empirical data from a small survey, why the exponential utility function may be too conservative. Bodie and Crane (1997) found by conducting a survey that the proportion of total assets held in equities declined with age and increased with wealth of the respondents. A more recent study by Faig (2002) conducted in Europe confirms a wide variety of investor behavior and that decreasing relative risk aversion can be shown as empirically valid. A different approach of modeling investor preference, called prospect theory, has been put forward by Kahnemann and Tversky (1979), postulating that investors are less concerned about their wealth as an absolute number than they are concerned about changes in wealth. They construct a utility function from a number of linear segments, each representing the utility of positive and negative changes in wealth.

5. A general approach to modeling utility

Following the empirical evidence about possible investor preference, we propose to model directly in the space of risk aversion rather than first defining a certain type of

utility function (which may or may not fit well) and then estimating its parameters. We have found an efficient way of representing the utility function as a piecewise exponential function with K pieces, where each piece represents a certain absolute risk aversion α_i, where $i = 1, \ldots, K$.

Let $\widehat{W}_i$, $i = 1, \ldots, K$, be discrete wealth levels representing the borders of each piece i, such that below each $\widehat{W}_i$ the risk aversion is α_i and above $\widehat{W}_i$ (till $\widehat{W}_{i+1}$) the risk aversion is α_{i+1}, for all $i = 1, \ldots, K$. For each piece i we represent utility using the exponential function. Thus, for $\widehat{W}_i \leqslant W \leqslant \widehat{W}_{i+1}$, $u_i(W_i) = a_i - b_i \exp(-\alpha_i W_i)$ and the first derivative with respect to wealth is $u_i'(W_i) = b_i \alpha_i \exp(-\alpha_i W_i)$. The absolute risk aversions α_i are computed in such a way that they represent the desired function of risk aversion versus wealth. We determine the coefficients of the exponential functions of each piece i by matching function values and first derivatives at the intersections $\widehat{W}_i$. Thus, at each wealth level $\widehat{W}_i$, representing the border between risk aversion α_i and α_{i+1}, we obtain the following two equations

$$a_i - b_i e^{-\alpha_i \widehat{W}_i} = a_{i+1} - b_{i+1} e^{-\alpha_{i+1} \widehat{W}_i},$$
$$b_i \alpha_i e^{-\alpha_i \widehat{W}_i} = b_{i+1} \alpha_{i+1} e^{-\alpha_{i+1} \widehat{W}_i},$$

from which we calculate the coefficients a_{i+1} and b_{i+1} as

$$b_{i+1} = b_i \frac{\alpha_i}{\alpha_{i+1}} e^{(\alpha_{i+1} - \alpha_1) \widehat{W}_i},$$
$$a_{i+1} = a_i - b_i \left(1 - \frac{\alpha_i}{\alpha_{i+1}}\right) e^{-\alpha_i \widehat{W}_i},$$

where we set arbitrarily $a_1 = 0$ and $b_1 = 1$.

The piecewise exponential function may span the whole range of attainable wealth levels. Starting from parameters $a_1 = 0$ and $b_1 = 1$ and given risk versions α_i, we compute all parameters a_{i+1} and b_{i+1} for each $i = 1, \ldots, K$.

To test the piecewise approximation, we could set the risk aversion α_i to represent typical utility functions, for example, setting each $\alpha_i = \alpha$ would result in constant absolute risk aversion (CARA) or setting $\alpha_i = \gamma / \widehat{W}_i$ would represent constant relative risk aversion (CRRA).

We are now in a position to set the piecewise absolute risk aversions to fit the risk aversion of the investor. Typically, we would model investors as having decreasing absolute risk aversion but either increasing or decreasing relative risk aversion. For example, the relative risk aversion may be increasing or decreasing from γ_0 at $\widehat{W}_0$, to γ_K at $\widehat{W}_K$. For a constant rate of change Δ in relative risk aversion, this could be modeled by setting

$$\gamma_{i+1} = \gamma_i + \Delta$$

where

$$\Delta = \frac{\gamma_K - \gamma_0}{\widehat{W}_K - \widehat{W}_0}$$

and

$$\alpha_i = \gamma_i / \widehat{W}_i.$$

In addition we may wish to represent constant relative risk aversion below $\widehat{W}_0$ and above $\widehat{W}_K$. In order to do so we append to the piecewise exponential representation on each side a CRRA piece represented by the power function $u(W) = c\frac{W^{1-\gamma}-1}{1-\gamma} + d$, $\gamma > 1$, with its first derivative $u'(W) = cW^{-\gamma}$. We then calculate the parameters of the power function as

$$c = \alpha e^{\alpha \widehat{W}} \widehat{W}^{\gamma},$$
$$d = -e^{\alpha \widehat{W}} - c\frac{\widehat{W}^{1-\gamma} - 1}{1-\gamma},$$

where in the formula for obtaining the parameters c_0 and d_0 for the lower CRRA piece we use $\alpha = \alpha_1$ and $\widehat{W} = \widehat{W}_0$ and for obtaining c_K and d_K for the upper CRRA piece we use $\alpha = \alpha_K$ and $\widehat{W} = \widehat{W}_K$. The formulas arise from setting at the intersections at $\widehat{W}_0$ and $\widehat{W}_K$ the function value and first derivative of the power piece equal to the function value and first derivative of the adjacent exponential piece, respectively.

As result we obtain a smooth, monotonically increasing and concave utility function that approximates arbitrarily closely (depending on the number of pieces used) the function of relative risk aversion representing the investor. Figure 1 displays the function of absolute risk tolerance versus wealth, for CARA and CRRA as well as two examples of representations of increasing and decreasing relative risk aversion.

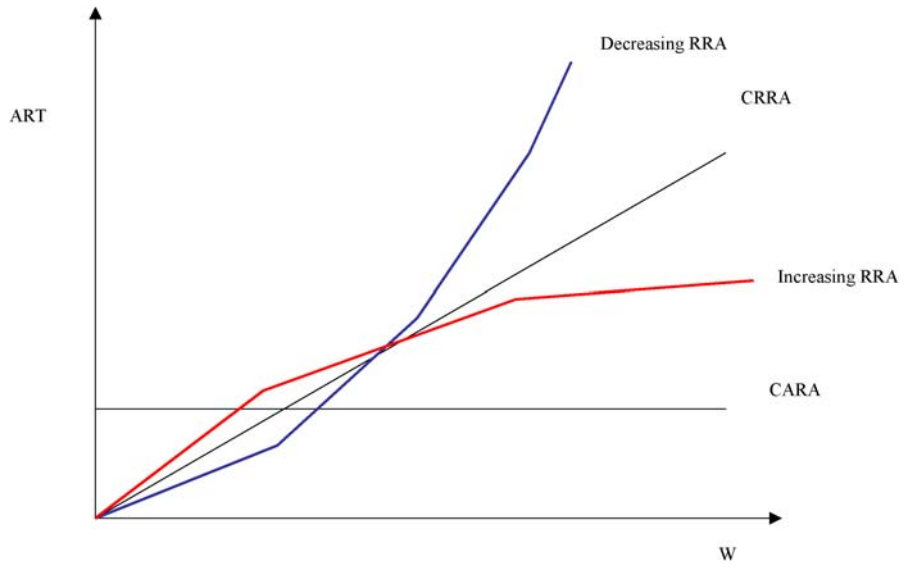


Fig. 1. Modeling risk aversion.

The piecewise exponential modeling represents a novel approach to representing an investor's risk aversion. We have found it easier to determine directly the risk aversion of an investor, e.g., by a questionnaire, as compared to determining the utility function by the certainty equivalent method or gain and loss equivalence method of comparing lotteries; see, e.g., Keeney and Raiffa (1976). Details of how to determine the risk aversion of an investor in this framework will be presented in a separate paper.

6. Dynamic portfolio choice

We now extend the single-period utility maximization model to a multi-period setting. Let $t = 0, \ldots, T$ be discrete time periods, with T the investment horizon. Let R_t be the random vector of asset returns in time periods t. Let $y_t = (y_1, \ldots, y_n)_t$ be the amount of money invested in the different asset classes $i = 1, \ldots, n$ at time t. Scalars W_0 and s_t, $t = 0, \ldots, T-1$, represent the initial wealth and possible cash flows (deposits positive and withdrawals negative) over time, respectively. The following problem states the multi-period investment problem:[1]

$$
\begin{aligned}
&\max Eu\big(e^T y_T\big),\\
&e^T y_0 = W_0 + s_0,\\
&-R_{t-1}^T y_{t-1} + e^T y_t = s_t, \quad t = 1, \ldots, T,\\
&y_t \geqslant 0, \quad W_0, s_0, \ldots, s_{T-1} \text{ given}, \ s_T = 0.
\end{aligned}
$$

At the beginning, the initial wealth plus the cash flow $(W_0 + s_0)$ are invested among the n asset classes. The returns of the investment plus the next cash flow are available for investment at the beginning of the next period. At the beginning of each period during the investment horizon, the investor can freely rearrange the portfolio. At the end of the investment horizon, the final wealth in the portfolio, W_T, is evaluated using the utility function $u(W_T)$. Short-selling and borrowing is ruled out, but could be easily introduced into the problem. Asset n could be a risk-free asset, but is treated like any other asset, since no distinction is necessary.

Instead of maximizing the utility of terminal wealth, we could maximize the discounted utilities of wealth in each period,

$$
\max \sum_{t=1}^{T} \delta^{-t} u_t\big(e^T y_t\big),
$$

where δ represents the discount factor. This concept of an additive discounted utility represents a straightforward extension and proved very useful in controlling, say, downside risk in every period. As an extension, the Kahnemenn and Tversky utility could be represented in such a way.

[1] T as a superscript always means transpose, while as a subscript T always denotes the last period of the investment horizon.

Defining x_t, $t = 0, T-1$, as the vector of fractions invested in each asset class in each period, we can write $x_t = y_t/(W_t + s_t)$, where we define the wealth available in each period (before adding cash flow) as W_t, $W_t = R_{t-1}x_{t-1}(W_{t-1} + s_{t-1})$. We can then write the model as

$$\begin{aligned}
&\max Eu(W_T),\\
&e^T x_t = 1, \quad t = 0, T-1,\\
&W_{t+1} = R_t x_t (W_t + s_t), \quad t = 0, \ldots, T-1,\\
&y_t \geqslant 0, \quad W_0, s_0, \ldots, s_{T-1} \text{ given}, \ s_T = 0.
\end{aligned}$$

Here one can see that for serially independent asset returns wealth is a single state connecting one period with the next. Now we write the problem as the dynamic programming recursion

$$\begin{aligned}
&u_t(W_t) = \max E\ u_{t+1}\big((W_t + s_t)R_t x_t\big),\\
&e^T x_t = 1,\\
&Ax_t = b, \quad l \leqslant x_t \leqslant u,
\end{aligned}$$

where $u_T(W_T) = u(W)$, $W_{t+1} = (W_t + s_t)R_t x_t$, W_0 given.

One can now see that the multi-period problem is composed of a series of single-period utility maximization problems, where for the last period the actual utility function is used and for all other periods the implied (through the dynamic programming algorithm) "utility-to-go" utility function is employed. Referring to the Samuelson and Merton result discussed above, for the CRRA utility function, all implied utility-to-go functions are also of the type CRRA, with the same risk aversion coefficient γ. This results in the aforementioned fixed-mix strategies.

6.1. Dynamic stochastic programming and Monte Carlo sampling

In practice, we need to resort to Monte Carlo sampling to estimate the expected utility of the single-period utility maximizing problem of each period. Let R_t^ω, $\omega \in S_t^i$, and R_t^ω, $\omega \in S_t^o$, $t = 1, \ldots, T-1$, be independent samples of the return distributions for each period t. The sample S_t^i includes the in-sample returns used for generating the single-period utility maximization problems and the sample S_t^o represents the out-of-sample return used for evaluating the obtained solution. Using two different samples, one for optimizing and the other for evaluating, prevents optimization bias. We represent the problem as

$$\begin{aligned}
&\hat{\hat{u}}_t(W_t) = \max \frac{1}{|S_t^i|} \sum_{\omega \in S_t^i} \hat{u}_{t+1}\big((W_t + s_t)R_t^\omega x_t\big),\\
&e^T x_t = 1,\\
&Ax_t = b, \quad l \leqslant x_t \leqslant u,
\end{aligned}$$

and we define $\hat{u}_t(W_t) = \frac{1}{|S_t^o|} \sum_{\omega \in S_t^o} \hat{u}_{t+1}((W_t + s_t) R_t^\omega x_t)$, where $\hat{u}_T(W_T) = U(W)$, W_0 given. We parameterize in W_t to carry out the dynamic programming recursion. Note $\hat{\hat{u}}(.)$ refers to the in-sample estimate, whereas $\hat{u}(.)$ represents the out-of-sample estimate of the utility-to-go function. Depending on the sample size used, the in-sample estimate $\hat{\hat{u}}(.)$ would include a significant amount of optimization bias that consequently would be carried forward between stages, whereas the out-of-sample estimate $\hat{u}(.)$ of the portfolio decision represents an independent evaluation without any optimization bias.

The dynamic optimization problem can now be solved using a backward dynamic programming recursion, conditioning on wealth. Starting at period $T-1$ we parameterize the wealth into K discrete wealth levels, W_{T-1}^k, $k = 1, \ldots, K$, and solve the period $T-1$ problem K times using sample S_{T-1}^i, and obtain solutions $\hat{x}_{T-1}^k$. We evaluate the obtained solutions by computing $\hat{u}_{T-1}^k = \frac{1}{|S_{T-1}^o|} \sum_{\omega \in S_{T-1}^o} \hat{u}_T((W_{T-1}^k + s_{T-1}) R_t^\omega x_t)$ and obtain for each parameterized value W_{T-1}^k a corresponding value of $\hat{u}_{T-1}^k$, which pairs represent K points of the Monte Carlo estimate of the value function $(u_{T-1}(W_{T-1}))$. We interpolate between those points, using an appropriate accurate interpolation technique, to obtain a smooth functional form. The value function $u_{T-1}(W_{T-1})$ in period $T-1$ is the induced utility function for the period $T-2$ single-period optimization, and we repeat the process until all optimizations in period 1 are done. Finally, in period 0, the initial wealth is known and we conduct the final optimization using the period 1 value function as implied utility function $u_1(W_1)$. In each period in the backward recursion, we use a different independent sample of large size for the evaluation: thus, the sampling error is small and cancels out over the different rebalancing periods. The sampling-resampling procedure is a crucial part of the solution algorithm, because it prevents the dynamic recursion from carrying forward and accumulating sampling bias when solving for a large number of periods.

6.2. Serially dependent asset returns

In the case of serial dependency of asset returns, we can extend the model and consider the return $R_t | R_{t-1}$ conditioned on the previous period return vector. For example, a vector autoregressive process (VAR(1)) of lag one would fit such a description. In this case we define

$$R_t = C + A R_{t-1} + \varepsilon,$$

where C is an intercept vector and A is an $n \times n$ matrix of coefficients obtained from n least-squares regressions. The problem is stated as

$$u_t(W_t, R_{t-1}) = \max E\, u_{t+1}\big((W_t + s_t) R_t | R_{t-1} x_t\big),$$
$$e^T x_t = 1,$$
$$A x_t = b, \quad l \leqslant x_t \leqslant u,$$

where $u_T(W_T) = u(W)$, $W_{t+1} = (W_t + s_t) R_t | R_{t-1} x_t$, W_0, R_{-1} given.

A general lag one vector autoregressive process with n asset classes requires $n + 1$ state variables and may exceed the limits of the stochastic dynamic programming approach. However, a restricted autoregressive process of lag one with a limited number of predicting variables may well be computationally tractable and statistically valid. Thus using, say, three predicting variables will lead to four state variables. The resulting dynamic program can still be solved accurately in reasonable time.

6.3. A fast method for normally distributed asset returns

For multivariate normally distributed asset returns, we can algorithmically take advantage of the fact that, for monotonically increasing concave utility functions, the optimal solution is mean-variance efficient. Instead of solving the recursive non-linear optimization problem, we can search a pre-computed set of mean-variance efficient solutions for the one maximizing the utility or value-to-go function. To maintain optimality in each stage, we need to ensure that the utility function, $u_t(W_t)$, induced from $\max E u_{t+1}((W_t + s_t)R_t x_t)$ given the constraints, is also a monotonically increasing concave function, as is the utility function $u(W)$ at the end of the investment horizon. The dynamic programming based proof of this theorem is omitted. Defining $\eta_{k,t}$ as the return distribution of the kth allocation point $(\mu_{k,t}, \sigma_{k,t})$ on the mean-variance efficient frontier in period t, i.e., $\eta_{k,t} = N(\mu_{k,t}, \sigma_{k,t})$, we can write the dynamic programming recursion as

$$u_t(W_t) = \max_k E\, u_{t+1}\big((W_t + s_t)\eta_{k,t}\big)$$

where $u_T(W_T) = u(W)$, $W_{t+1,k} = (W_t + s_t)\eta_{k,t}$, W_0 given.

An effective search is used to speed-up the optimization by avoiding having to evaluate all distributions $\eta_{k,t}$. The work required for this recursion is independent of the number of assets and, like the general recursion above, linear in the number of stages.

This fast approach is very well suited for solving the problem with a restricted autoregressive return process, where the error terms of the restricted vector autoregression are assumed distributed as multivariate normal. Details of this approach will be discussed in a separate paper.

7. Numerical results

7.1. Data assumptions

For estimating the asset return distributions, we used historical monthly time series from January 1974 to December 2004 for stocks and from January 1986 to December 2004 for bonds, based on the data available from the Datastream Access database. Stocks were classified into US stocks (represented by the MSCI US index) and international stocks (represented by the MSCI EAFE&C index). Bonds were distinguished into US

Corporate Bonds (represented by the Salomon Brothers US Corp Bnd index) and US Gvnt Bonds (represented by the GP Morgan US Gov Bnd index). For cash we used 3 month Treasury notes (represented by the GP Morgan US 3M index). For obtaining the i.i.d. distributions, we estimated the correlation matrix using directly the historical data points for each time series from January 1986 to December 2004. We estimated separate correlation matrices for representing normally and lognormally distributed returns. The estimates for the correlation matrices and the means and standard deviations are presented in Table 2. However, for corporate bonds we lowered the historical expectation of 9.49% to 9.0% per annum, reflecting a more appropriate number, since expected corporate bond returns appeared as too high during this period and were overly dominating stocks.

Table 2
Data estimates for normal and log normal distributions

Historical means and standard deviations					
	US stocks	Int stocks	Corp bnd	Gvnt bnd	Cash
Mean	10.80	10.37	9.49	7.90	5.61
STD	15.72	16.75	6.57	4.89	0.70
Regression-based means and standard deviations					
	US stocks	Int stocks	Corp bnd	Gvnt bnd	Cash
Mean	14.1	12.5	9.24	7.92	5.77
STD	15.72	16.75	6.57	4.89	0.70
Correlation matrix for normal model					
	US stocks	Int stocks	Corp bnd	Gvnt bnd	Cash
US stocks	1.00	0.601	0.247	0.062	0.094
Int stocks	0.601	1.00	0.125	0.027	0.006
Corp bnd	0.247	0.125	1.00	0.883	0.194
Gvnt bnd	0.062	0.027	0.883	1.00	0.27
Cash	0.094	0.006	0.194	0.27	1.00
Correlation matrix for log-normal model					
	US stocks	Int stocks	Corp bnd	Gvnt bnd	Cash
US stocks	1.00	0.609	0.236	0.05	0.083
Int stocks	0.609	1.00	0.124	0.02	−0.002
Corp bnd	0.236	0.124	1.00	0.884	0.195
Gvnt bnd	0.05	0.02	0.884	1.00	0.271
Cash	0.083	−0.002	0.195	0.271	1.00

In addition to using the historical mean returns directly, we also estimated a constant growth rate for each asset class. By fitting a constant growth rate to the data we accounted for the dips at the beginning and end of the time series of stock returns. Using a linear regression on the $1 + \ln(R^C_{i,t})$ compounded returns, the least square problem

$$\min\bigl(a_i + b_i t - 1 - \ln\bigl(R^C_{i,t}\bigr)\bigr)^2$$

results in an intercept a_i and a coefficient $b_i = (1 + \ln(\bar{r}_i))$, from which the growth rate of $\bar{r}_i$ for each asset class i can be obtained. (The latter could be used as forward-looking unconditional expected returns.) The standard deviations (vector σ) were estimated directly from the monthly data. The historical time series in logarithmic representation as well as the linear function of estimated constant growth from the regression are represented in Figure 2 for US stocks, international stocks, and (for comparison) all stocks, represented by the MSCI World total return index, and in Figure 3 for corporate bonds, government bonds and cash.

Assuming either normal or lognormal asset returns distributions does not take into account properly the higher moments of the joint return distributions of the asset classes. We therefore used a bootstrapping procedure to capture the true co-movements of the asset classes. In order to do so, we subtracted from each historical time series its historical mean and divided by its historical standard deviation to obtain the vector of residuals ε_t. With $\bar{r}^H$ and σ^H representing the historical vectors of means and standard deviations, respectively, and $\varepsilon_t = (R_t - \bar{r}^H)/\sigma^H$, the returns distribution is represented as $R_t = \bar{r} + \sigma\varepsilon_t$. For generating one yearly return observation via the bootstrapping scheme, we sampled 12 independent observations with replacement from the historical monthly residuals, and combined them into a corresponding yearly return by compounding, multiplying by the standard deviation, and adding the (forward looking unconditional) mean. For the discussion below we used the more conservative historical mean estimates, rather than the one obtained by the regression, and the variances were estimated directly from the monthly historical time series.

We remark that estimating the means of stock returns, given the data at hand, may not be very accurate. For example, our estimates for the annual means based on 374 monthly historical data have an estimated standard deviation of about 2.8%. Estimates of standard deviations and correlations are more accurate. The regression procedure provides a good way to estimate means. The regression results for stocks, while larger than the historical estimates, are within the confidence interval obtained for the historical mean estimates.

For the effect of estimation errors on optimal portfolio choice see, e.g., Chopra and Ziemba (1993) and on turnover Chopra (1993), showing that estimation errors in the means have a significantly larger effect than estimation errors in the variances and covariances. Simulating estimation error by adding zero mean i.i.d normal distributions to the data, and comparing certainty equivalent cash (wealth) in a single-period mean-variance optimization of 10 randomly selected stocks of the Dow Jones industrial average, for a risk tolerance of 50, errors in the mean estimate resulted in a 11 times larger loss of certainty equivalent cash as errors in the variances, and errors in

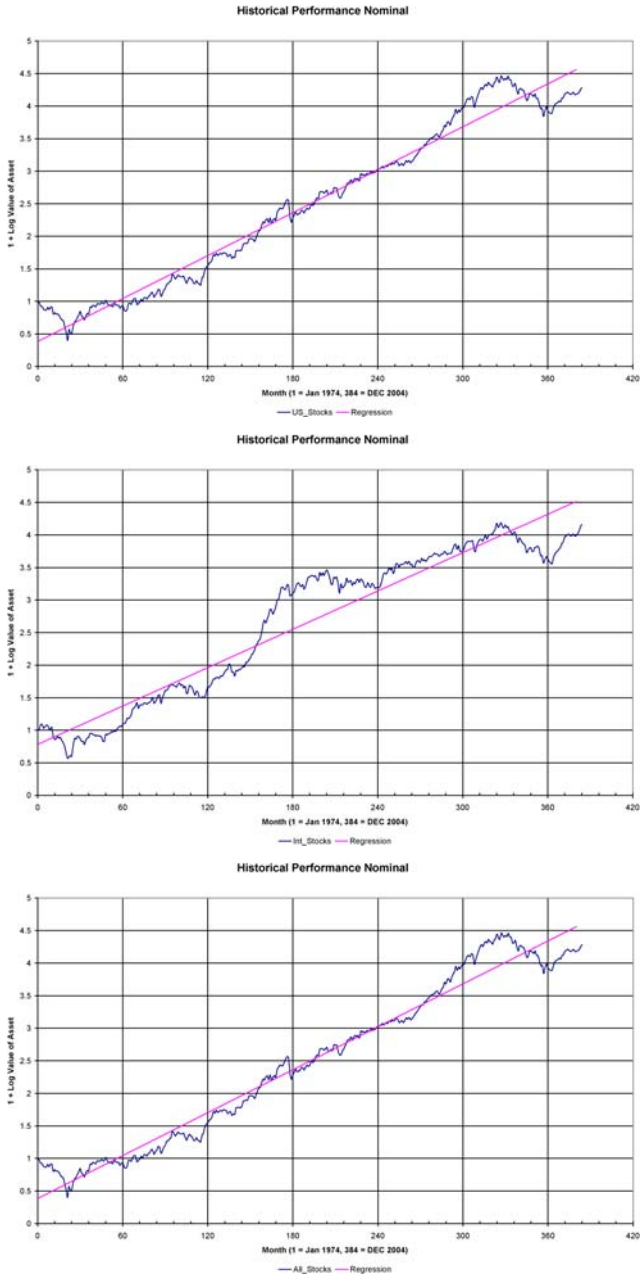


Fig. 2. Monthly stock performance, January 1974 – December 2004 (US stocks, international stocks, whole world).

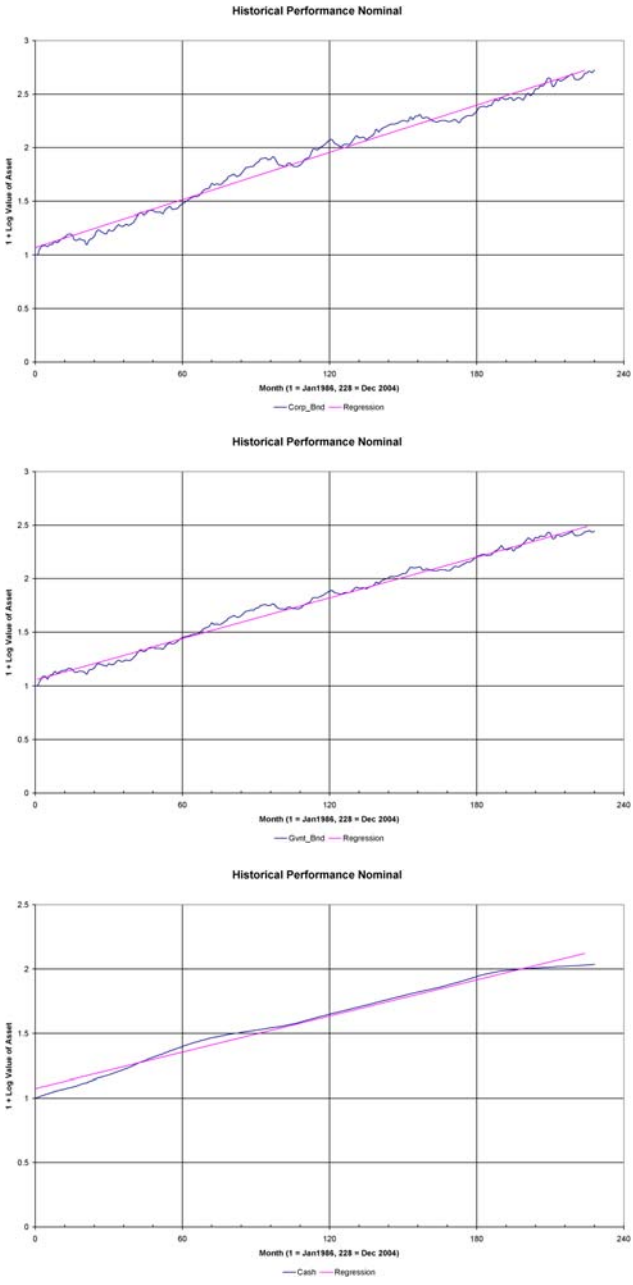


Fig. 3. Monthly bond performance, January 1986 – December 2004 (corporate bonds, government bonds, cash).

the variances resulted in 2 times larger loss than errors in the covariances. The effects were shown to increase with increasing risk tolerance and with increasing magnitude of errors. Already earlier, Kallberg and Ziemba (1981, 1984) concluded that errors in the mean estimate matter most. The effect of errors in the estimates of means relative to variances and covariances on certainty equivalent wealth obviously must increase with larger risk tolerance, since at a risk tolerance of zero, where only variances and covariances enter the optimization problem, exclusively errors in variances and covariances matter, and at an infinite risk tolerance, where only means enter the optimization problem, exclusively errors in mean estimates have an influence. Michaud (1989) showed that noisy forecasts in mean-variance optimization may lead to suboptimal portfolios, where assets with positive estimation error in the mean forecasts are significantly overweighted and assets with negative estimation error are significantly under-weighted in the "optimal" portfolio, and proposed an approach based on re-sampling (bootstrapping) to counteract this effect. As a practical approach to counter estimation error, Connor (1997) proposed for linear regression based forecasting models to use Bayesian adjusted priors to shrink mean estimates with large observed estimation errors.

A different approach for obtaining unconditional means is to use an assumed mean-variance efficient portfolio, e.g., a broad index, and to infer the unconditional means from the standard deviations of returns, by viewing the portfolio as the optimal solution of a corresponding mean-variance problem with an appropriate risk aversion coefficient. This procedure of Grinold (1999) is called "grapes from wine", and represents an efficient way to calibrate mean returns. Besides estimating standard deviations based on historical time series, also implied volatilities based on observed option prices could be used.

7.2. An investment example

In order to demonstrate dynamic investment policies obtained from our dynamic portfolio choice model, we discuss as an illustrative example a very typical investment situation. An investor has a current wealth of \$100k and plans to contribute \$15k per year for the next 20 years. What is the distribution of wealth at the end of the investment horizon, reflecting various reasonable assumptions about the investor's risk aversion profile? We discuss four cases: (A) using the CARA utility function, (B) using increasing relative risk aversion but decreasing absolute risk aversion, (C) using decreasing relative risk aversion and decreasing absolute risk aversion, and (D) using a quadratic penalty of under-performing a target.

For the exponential utility function (case A), we assumed an absolute risk aversion coefficient of $\lambda = 2$. Figure 4 presents the optimal asset allocation (as a function of wealth) at various times, e.g., for one year (top), 10 years (middle) and 19 years (bottom) to go. The optimal asset allocation is not constant but varies significantly with the investment horizon and the wealth already accumulated. The case of one year to go is special because it shows the results for different wealth levels of a single-period optimization with the investor's original utility function. All other period allocations

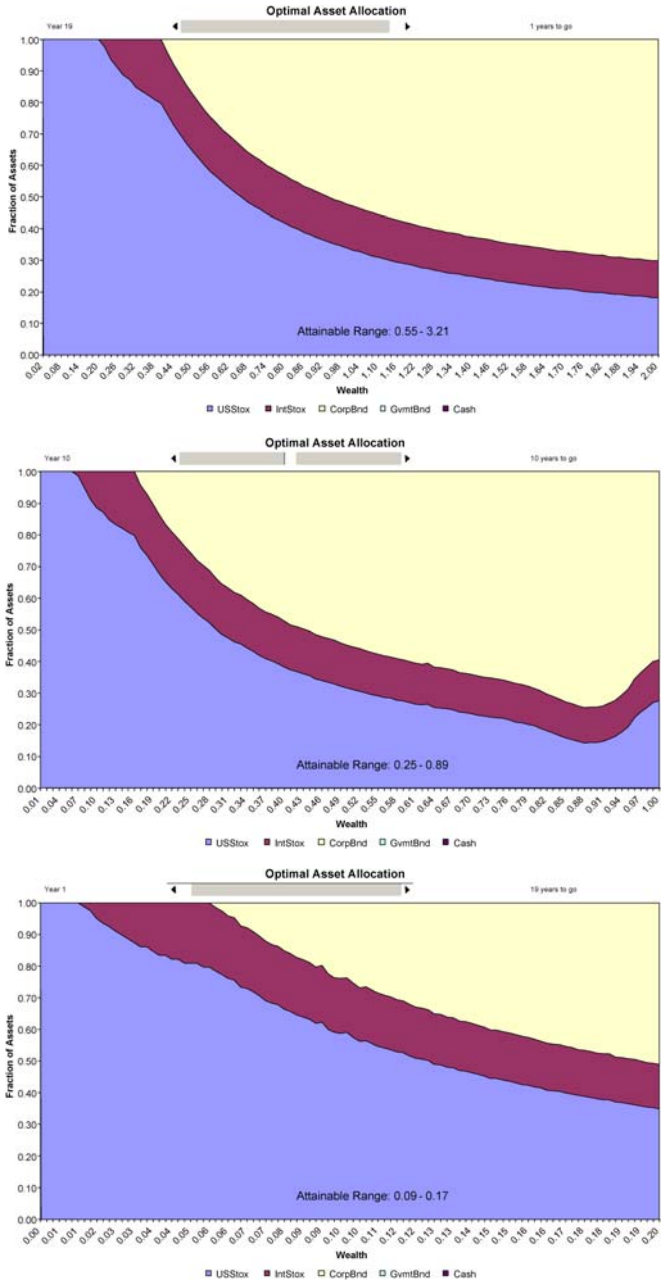


Fig. 4. Investment example: initial \$100k plus \$15k per year, 20-year horizon: exponential utility, investment strategy.

represent the result for using the implied utility function from the dynamic programming algorithm. In the last rebalancing period, the more money that is in the investor's account the less risk he/she takes, and accordingly the fraction of stocks decrease with increasing account value. When the account is under-performing a certain wealth level the optimal strategy prescribes to invest entirely in stocks, where the amount to be put internationally versus domestically varies with account value. This may be a significant strain on the investor, but the problem could be corrected by using a constraint restricting the maximum exposure to stocks. If there was no borrowing constraint, we know from analytical solutions for HARA policies that the optimal investment would prescribe for low wealth levels to borrow funds and use this money to buy more than 100% stocks, thus attempting to leverage the funds available. However, this would not be practical in a low-wealth situation, since such a strategy would quickly exceed margin requirements and thus would not be implementable. This demonstrates the importance of considering a borrowing constraint as part of the investment problem.

The dependence of the optimal asset allocation on wealth changes every year, as one can see by comparing the optimal asset allocation for only one year to go with the ones for 10 years to go and for 19 years to go. For very high wealth levels and long remaining horizons (e.g., 19 years to go) a small amount of cash enters the optimal portfolio. Figure 4 for the optimal asset allocation strategy also displays the attainable wealth range, obtained by simulation. The attainable range is defined by the wealth that is exceeded (left) or not exceeded (right) with 99.9% probability, respectively. For example, in year 19 the attainable range is between $0.55 and $3.21 million, in year 10 between $0.25 and $0.89 million, and in year 1 (after the first year of investment) between $0.09 and $0.17 million, including the cash flow at the beginning of the period.

Figure 5 (top) represents the expected value strategy, i.e., the optimal investment strategy implemented over time, assuming every year that the expected returns would be realized. At (the current) period zero, the optimal portfolio is 54% US stocks, 17% international stocks, and 29% corporate bonds. Cash and government bonds are not part of the initial portfolio. One can observe that from the optimal allocation at the outset of approximately 71% stocks and 29% bonds, the allocation changes gradually to approximately 37% stocks and 63% bonds at year 19, when the last portfolio decision is to be made. This path is similar to strategies that investment firms tend to prescribe their clients and recently have been implemented as life-cycle funds. The reasoning behind such strategies is to reduce stock allocation as the investment horizon shortens to prevent a possible significant loss without a good prospect of recovery before the end of the investment horizon. However, the strategies of practitioners and life-cycle funds are different to our dynamic strategy in that they do not react to prior performance. We can also view the dynamic strategy by starting from the expected value path, where stock allocation and risk is reduced as the remaining investment horizon shortens. In each period the stock allocation and the risk is reduced if the performance was better than expected (and thus the available wealth is larger), and the stock allocation and the risk is increased if the performance was worse than expected (and thus the wealth is smaller).

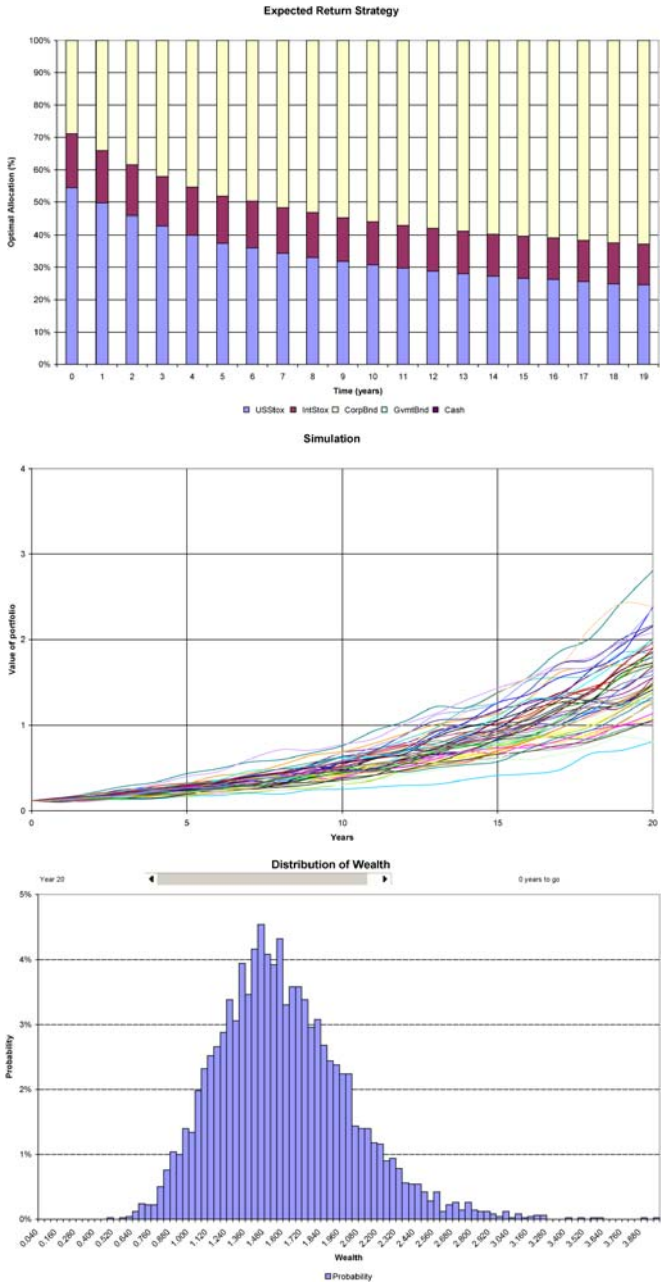


Fig. 5. Investment example: initial \$100k plus \$15k per year, 20-year horizon: exponential utility, simulations.

Table 3
Out-of-sample simulation results

	CEW	Mean	Std	99%	95%
Exponential RA = 2	1.412	1.564	0.424	0.770	0.943
Increasing RRA	1.440	1.575	0.452	0.771	0.937
Decreasing RRA	1.339	1.498	0.436	0.865	0.998
Quadratic	0.982	1.339	0.347	0.911	1.006
Downside	1.349	1.481	0.352	0.865	0.997

Figure 5 depicts (in the middle) the out-of-sample simulations of wealth over time and (at the bottom) the resulting marginal distribution of wealth at the end of the investment horizon. Table 3 summarizes the statistics of wealth obtained at the end of the investment horizon. The mean wealth is \$1.564 million, with a standard deviation of \$0.424 million. With 99% probability a wealth of \$770,000 is exceeded and with 95% probability a wealth of \$943,000. The certainty equivalent wealth is \$1.412 million.

Alleged shortcomings of the exponential utility function (constant absolute risk aversion) include that very high stock allocations at very low wealth levels may lead to a too-high risk burden on the investor, and a too-small relative risk aversion at very high wealth levels may lead to overly conservative asset allocations. While both shortcomings can be compensated for with lower and upper bounds on the stock allocation, another way is to model the proper risk aversion of the investor directly.

For the increasing relative risk aversion utility function (case B), we assumed that at a wealth below $W_L = 0.25$, the relative risk aversion is $\gamma = 2$, increasing linearly with wealth to $\gamma = 3.5$ at a wealth of $W_U = 3.5$ and then remaining at that level for larger wealth levels. This profile of relative risk aversion was modeled using the piecewise exponential utility representation discussed above, using 200 exponential (CARA) pieces between W_L and W_U, and CRRA pieces for wealth levels below W_L and above W_U. Besides increasing relative risk aversion, the function exhibits decreasing absolute risk aversion, since the absolute risk aversion decreases from 8 at $W_L = 0.25$ to 1 at $W_U = 3.5$, and the adjacent CRRA pieces also represent decreasing absolute risk aversion.

The results for increasing relative risk aversion are similar to the ones for the exponential case, but do not display the aforementioned shortcomings of the exponential utility function. The allocation of stocks at low levels of wealth are reasonable, with a maximum at about 64%, and the stock allocation at very high wealth levels does not become as conservative as in the exponential case. Again, the optimal strategies depend both on wealth level and remaining investment horizon, as Figure 6 presents for one year (top), ten years (middle) and 19 years (bottom) to go. The assumption of a relative risk aversion bounded between 2 and 3.5 is reflected in the flat response to changes in wealth levels. Figure 6 also presents the attainable wealth in year 19 between \$0.59 and \$3.44 million, in year 10 between \$0.27 and \$0.90 million, and in year 1 between \$0.10 and \$0.16 million.

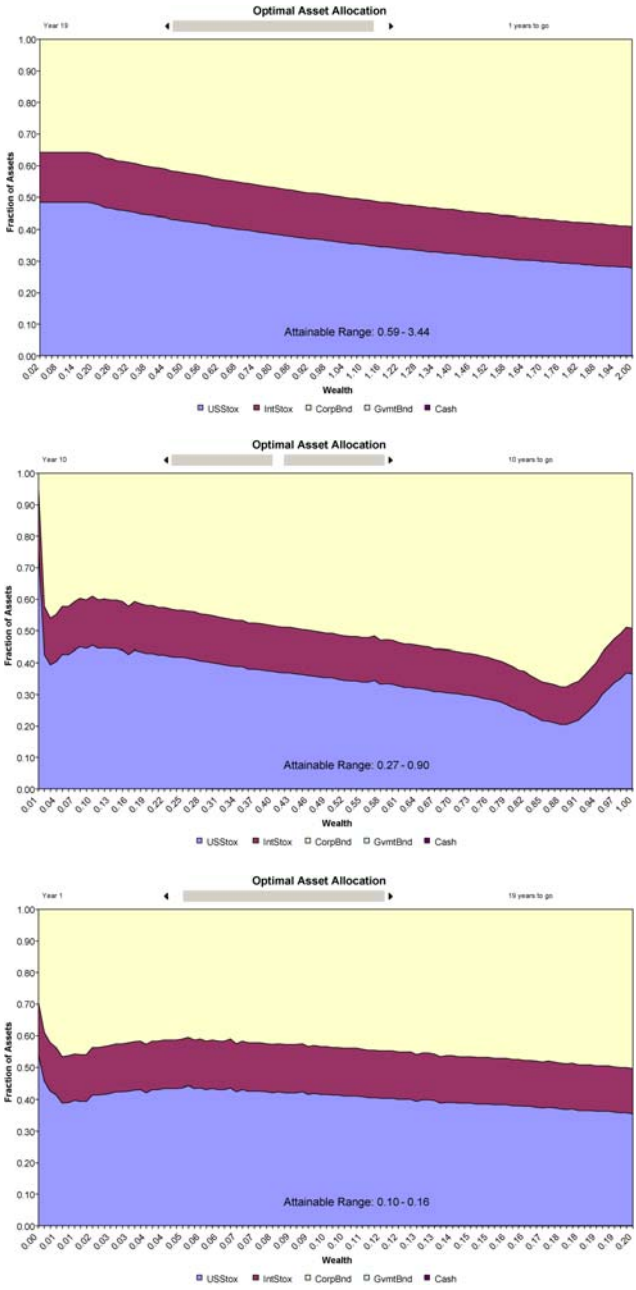


Fig. 6. Investment example: initial \$100k plus \$15k per year, 20-year horizon: increasing relative risk aversion, investment strategy.

Figure 7 represents the expected value strategy (top), the out-of-sample simulation of wealth (middle), and the marginal wealth distribution (bottom). The initial investment is 41% US stocks, 15% international stocks, and 44% corporate bonds. As in the exponential utility case above, government bonds and cash are not in the optimal initial portfolio. The expected value strategy calls for an initial optimal stock allocation of about 56%, to be decreased gradually to about 46% the last decision period. Again, we can describe the dynamic strategy in terms of deviation from the expected value strategy. The optimal dynamic strategy calls for a larger stock allocation and more risk compared to the expected value strategy in the case of under-performance (and therefore lower wealth) and smaller stock allocations and less risk in the case of out performance (and therefore larger wealth). The maximum and minimum stock allocation and the associated risk of the portfolio are within reasonable limits, which makes the increasing relative risk aversion strategy very practical for investors wishing to become more conservative at higher levels of wealth.

Not surprisingly, the out-of-sample results in Figure 7 (middle) are very similar to the exponential case in Figure 5 (middle) before. For the increasing relative risk aversion strategy, the mean wealth obtained at the end of the investment horizon is \$1.575 million with a standard deviation of \$0.452 million. With 99% probability a wealth of \$771,000 is exceeded and with 95% probability a wealth of \$937,000. The certainty equivalent wealth is \$1.44 million. The similar return results compared to the exponential utility case, combined with and the lack of undue risk exposure, underlines the practicability of the increasing relative risk aversion strategy.

Next, as case (C), we discuss decreasing relative risk aversion, which implies also decreasing absolute risk aversion. We assume that the relative risk aversion is high at $\gamma = 8.0$ below a wealth of $W_L = 1.0$ and decreases linearly to a low relative risk aversion of $\gamma = 1.01$ for wealth levels of $W_H = 1.5$ and above. An investor exhibiting this utility function becomes increasingly less risk averse with increasing wealth, but is very risk averse at low levels of wealth. This type of utility function is for investors who want to hedge the downside, e.g., wish not to under-perform a wealth of about \$1 million at the end of the investment horizon. Figure 8 shows the optimal strategy for this utility function for one (top), 10 (middle) and 19 (bottom) years to go. The optimal asset allocation in the last decision period (one period to go) reflects the risk aversion profile: at small levels of wealth the stock allocation is small at about 21% and increases to about 87% stocks at higher levels of wealth. Again, the asset allocation changes with the remaining investment horizon and wealth, where the point of low stock allocations shifts towards higher wealth levels as the remaining investment horizon decreases and the change of allocation becomes less gradual. We also observe very conservative investments at very low wealth levels, but these are out of the range of wealth that can be reasonably obtained. The initial portfolio is 11% US stocks, 10% international stocks, 60% corporate bonds, and 19% government bonds. Cash is not part of the initial portfolio. The expected value strategy in Figure 9 (top) shows that the stock allocation and thus the risk increases with time starting from about 21% stocks at the initial investment up to 50% stocks at the last decision period in year 19. This reflects the investor's

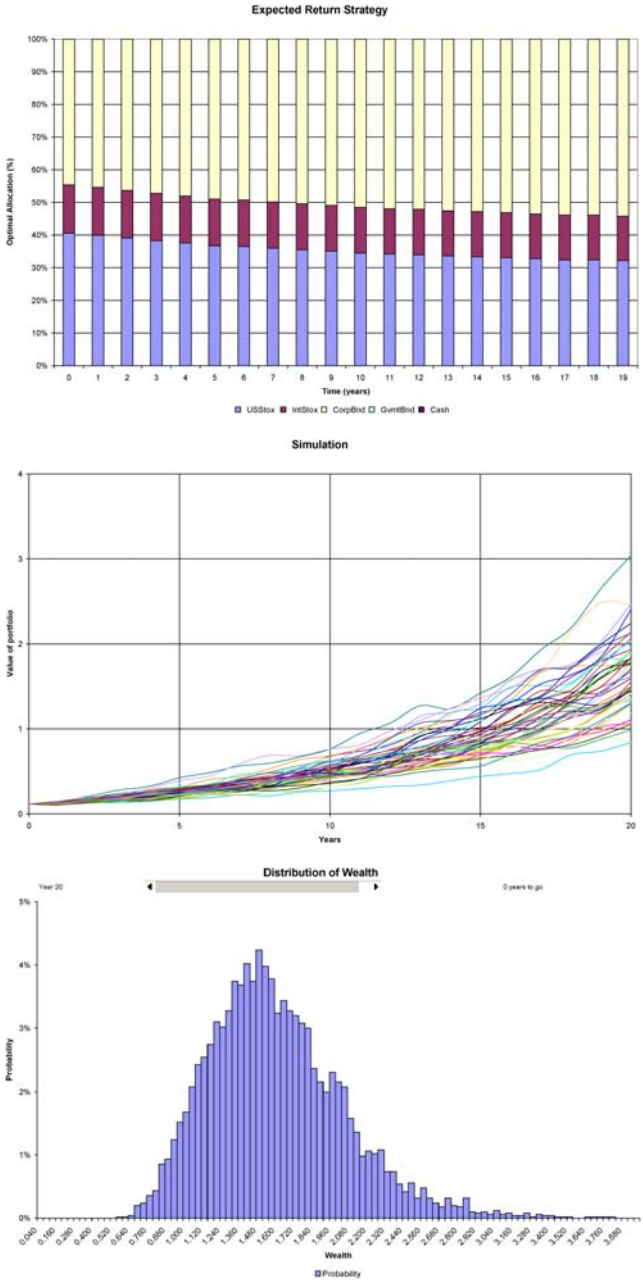


Fig. 7. Investment example: initial \$100k plus \$15k per year, 20-year horizon: increasing relative risk aversion, simulations.

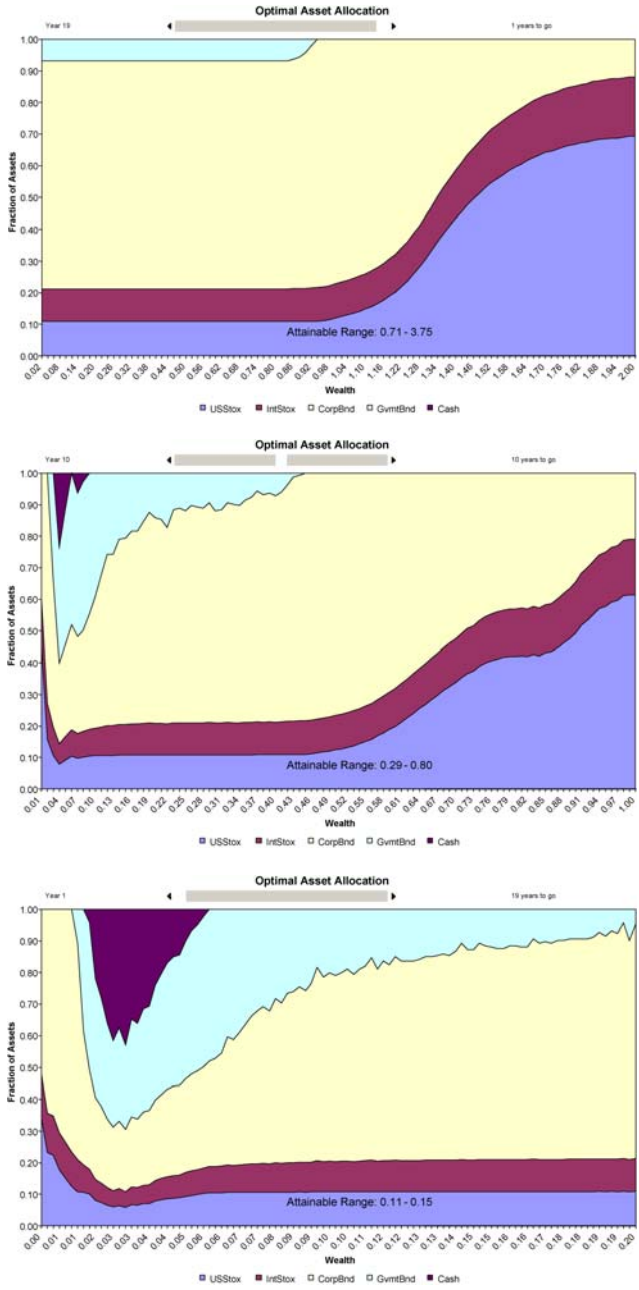


Fig. 8. Investment example: initial \$100k plus \$15k per year, 20-year horizon: decreasing relative risk aversion, investment strategy.

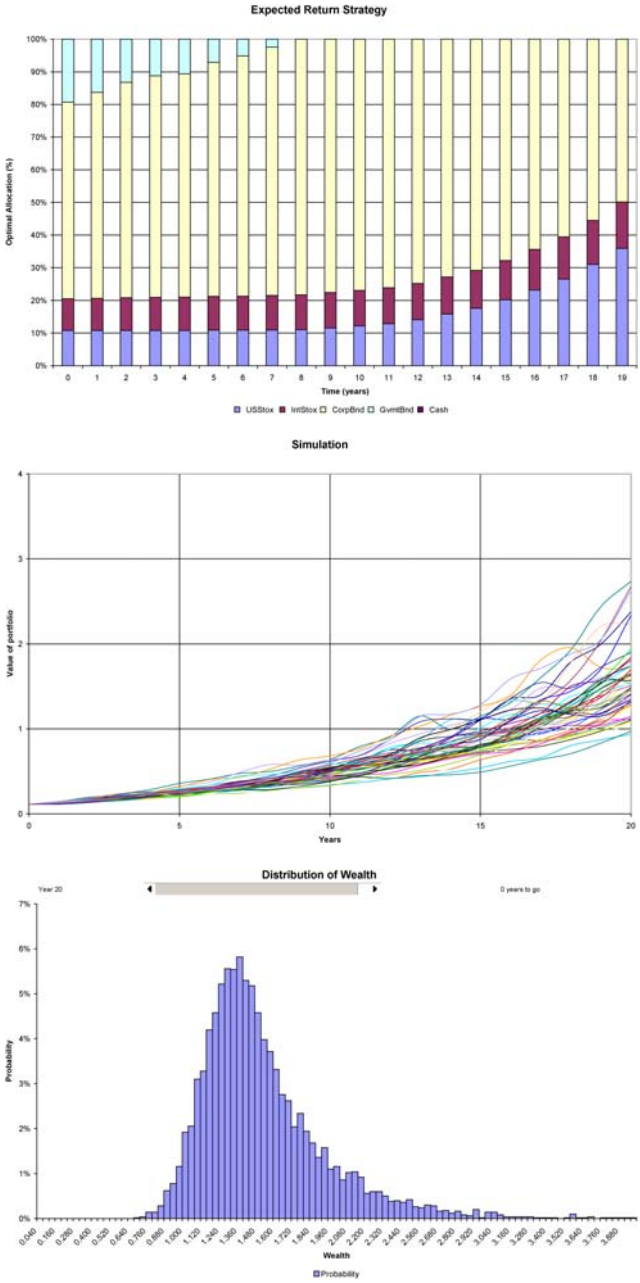


Fig. 9. Investment example: initial \$100k plus \$15k per year, 20-year horizon: decreasing relative risk aversion, simulations.

profile as a time path, where he/she is careful at low levels of wealth, but becomes increasingly less risk averse as wealth grows over time. Relative to the expected value path, the strategy prescribes in each period to increase the stock allocation and the risk if the performance was better than expected (and more wealth than expected is available) and to reduce the stock allocation and risk if under-performance occurred (and less wealth than expected is available). Figure 8 presents the attainable wealth in year 19 between \$0.71 and \$3.75 million, in year 10 between \$ 0.29 and \$0.80 million, and in year 1 between \$0.11 and \$0.15 million.

Figure 9 (in the middle) gives the out-of-sample simulation results, and (at the bottom) the marginal distribution of terminal wealth. In both views, one can observe that the downside is more protected than in the previous cases. The mean wealth obtained at the end of the investment horizon is \$1.498 million with a standard deviation of \$0.436 million. With 99% probability a wealth of \$866,000 is exceeded and with 95% probability a wealth of \$998,000. These are reasonable out-of-sample results for an investor wishing to protect his/her downside below \$1 million. The certainty equivalent wealth is \$1.339 million (Table 3).

The optimal dynamic strategies for increasing and decreasing relative risk aversion behave in a mirrored way. Assuming a crash in the stock market occurred, we can deduce that an investor with increasing relative risk aversion would react by increasing the stock allocation in order to make up for the loss, while the investor with decreasing relative risk aversion would reduce his/her stock allocation in order to further protect the downside. And we know that an investor with constant relative risk aversion (CRRA) would re-balance after the crash back to his/her original asset allocation. We may use this theoretical behavior to help infer the type of utility function that is most appropriate for an investor.

Case (D) reflects risk as a quadratic penalty of under-performing a target wealth. We assumed a target wealth of \$1M and traded off risk and expected return using a risk aversion of $\lambda_q/2 = 1000$. This is an extreme case reflecting an investor wishing to obtain a wealth of \$1 million very badly and therefore being prepared to forsake a significant part of the upside. Figure 10 presents the dynamic investment strategy for one (top), ten (middle) and 19 (bottom) years to go. Looking at the last rebalancing period, the optimal strategy reflects an increasing stock allocation for higher levels of wealth, starting slightly below the target wealth. This "critical" wealth level is where the target can still be reached with very high probability. Reducing wealth from large values, the closer we are to this critical wealth level, the more conservative the investment becomes, up to putting almost the entire portfolio into cash. For wealth levels above the target, stock allocations increase with increasing wealth up to 100%. For wealth levels below the critical level, the strategy becomes more risky with decreasing wealth, with stock allocations also rising up to 100%. This behavior reflects exactly the risk aversion represented by the quadratic downside utility function. For wealth levels above the target, the linear term (representing zero risk aversion) is dominant and leads to increased stock allocations up to 100 percent. For wealth levels below the critical point, the quadratic part of the utility function is dominant. Penalizing under-performance quadratically leads to

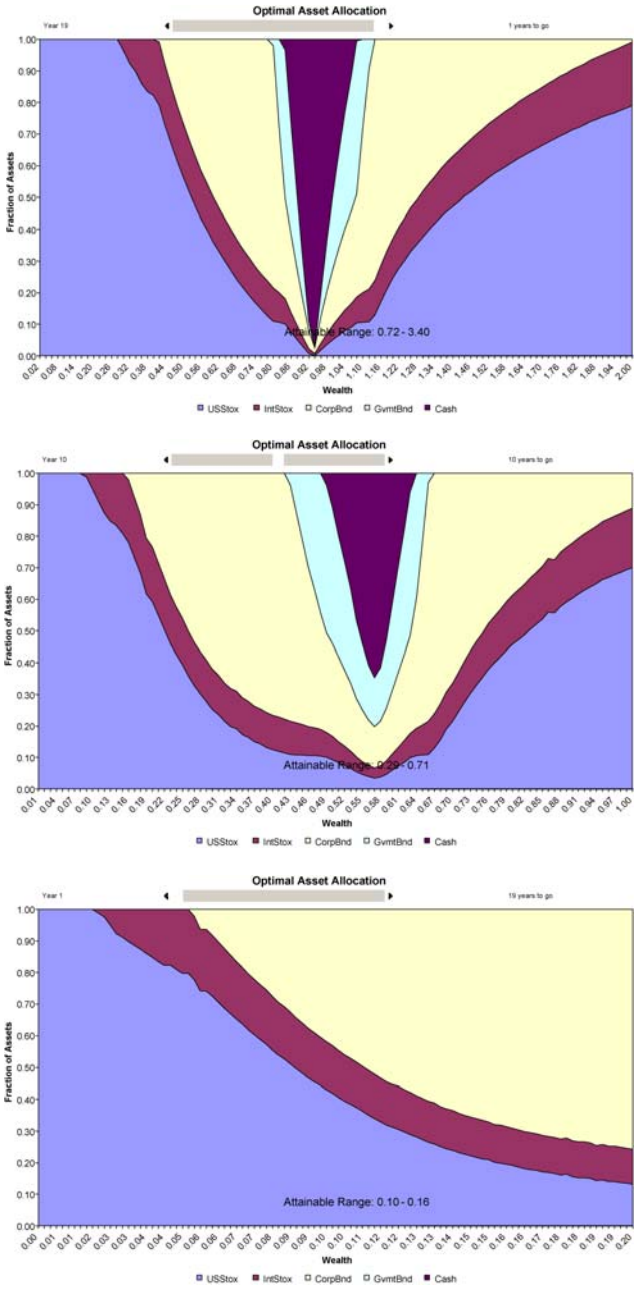


Fig. 10. Investment example: initial $100k plus $15k per year, 20-year horizon: quadratic utility, investment strategy.

a relative risk aversion that decreases with larger under-performance and thus increases with larger wealth. Therefore, the quadratic utility function reflects increasing relative risk aversion below the critical wealth and decreasing relative risk aversion above. With longer remaining investment horizons the critical point of reaching the target shifts to the left, and the allocation at that point becomes less conservative. With ten years to go, the critical wealth level is at about $0.58 million and the allocation is at about 65% cash. Interestingly, with 19 years to go, the reachable wealth levels are all below the critical point and the strategy is entirely one of decreasing relative risk aversion, calling for stock allocations from about 60% at low levels of wealth following a bad first year to 30% after an out-performing first year. Figure 10 represents the attainable wealth in year 19 between $0.72 and $3.40 million, in year 10 between $0.29 and $0.71 million, and in year 1 between $0.10 and $0.16 million.

The initial portfolio is 36% US stocks, 14% international stocks, and 50% corporate bonds. Government bonds and cash are not part of the initial optimal portfolio. The expected return strategy in Figure 11 (top) confirms this result. Starting at an initial stock allocation of 50%, the stock allocation is reduced gradually through year 13 and then increased from year 14 on. Given the savings rate and the results from the investments, the investor first starts in an under-performing state, then achieves the critical wealth level during year 14, and ends above the critical wealth.

The out-of-sample simulation results in Figure 11 (middle) show good downside protection and the marginal probability chart (bottom) is very steep on the left reflecting the downside protection the quadratic downside risk function is said to provide. However, even the assumption of a very large risk aversion coefficient does not lead to a zero probability of wealth below $1 million. The mean wealth obtained at the end of the investment horizon is $1.339 million with a standard deviation of $0.347 million. With 99% probability a wealth of $911,000 is exceeded and with 95% probability a wealth of $1.006 million. Thus, the target wealth of $1 million is exceeded with larger than 95% probability. The statistics in Table 3 give a certainty equivalent wealth of $998,000 as the lowest of all four utility functions, reflecting again the low emphasis on the upside displayed by the quadratic downside utility function.

7.3. The performance of dynamic strategies

Next we compare dynamic strategies obtained from using the different utility functions with a number of fixed-mix strategies, i.e., (1) US stocks only, (2) cash only, (3) all asset classes equally weighted, (4) risk averse, (5) medium, and (6) risk prone. Except for case (3), all fixed-mix strategies are mean-variance efficient and obtained from solving the single-period Markowitz problem. Figure 12 presents the efficient frontier and the efficient fixed-mix portfolios. We intentionally use the five representative mean-variance optimal fixed-mix strategies for comparison, because such portfolios are commonly held in practice. Many aggressive investors hold stocks-only portfolios, and many very conservative investors keep their funds entirely in money market accounts. Investment firms usually offer fund-of-funds portfolios, such as averse (often called

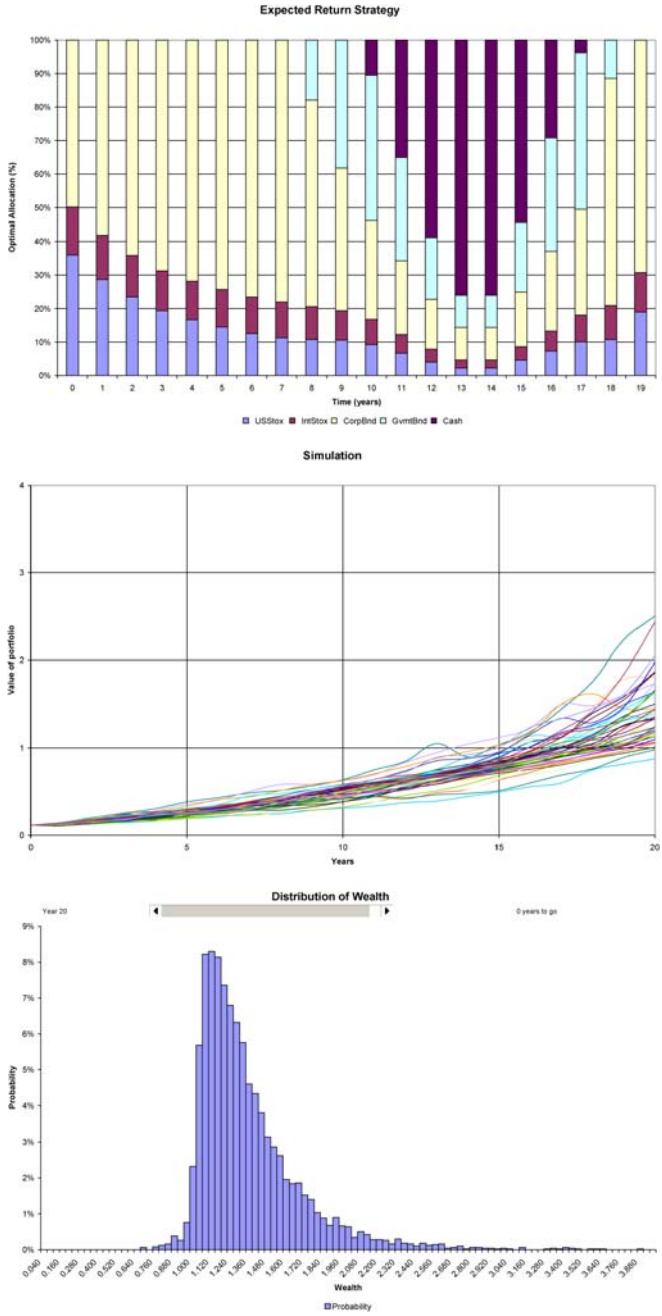


Fig. 11. Investment example: initial $100k plus $15k per year, 20-year horizon: quadratic utility, simulations.

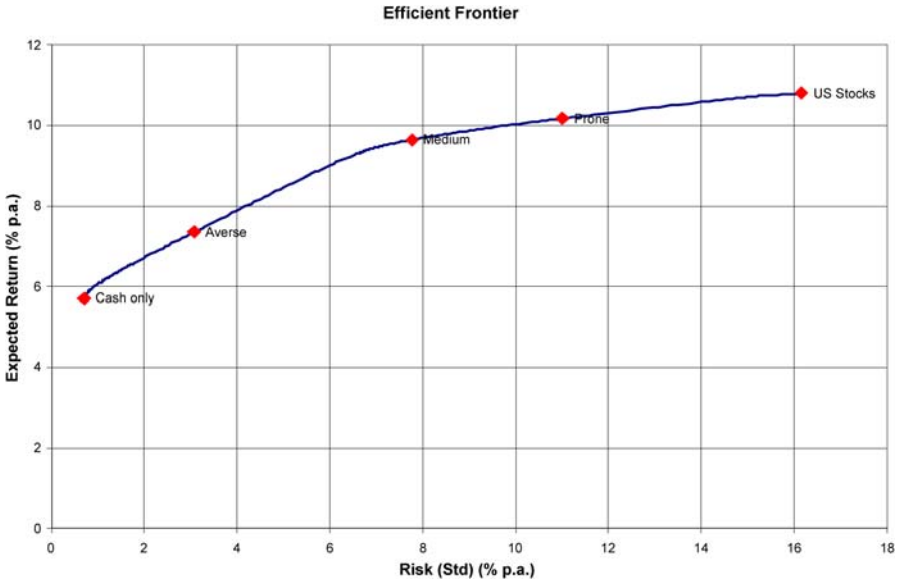


Fig. 12. Efficient frontier and efficient portfolios for fixed-mix strategies.

Table 4
Out-of-sample simulation results for various fixed-mix strategies

	Mean	Std	99%	95%
US stocks	1.825	1.065	0.469	0.660
Cash	0.868	0.019	0.822	0.834
Equally weighted	1.349	0.301	0.799	0.920
Averse	1.098	0.110	0.869	0.930
Medium	1.538	0.407	0.825	0.975
Prone	1.663	0.639	0.677	0.852

conservative), medium (often referred to as dynamic), and prone (often referred to as aggressive), which are sold to investors allegedly according to their risk profile. The equally weighted strategy represents a non-efficient portfolio for comparison.

Table 4 represents the out-of-sample simulation results for the various fixed-mix strategies. Obviously, the largest expected wealth of $1.825 million is obtained by the US Stocks portfolio, and the smallest expected wealth of $0.868 million by the cash-only portfolio. More interestingly, the cash-only portfolio exceeds a wealth of $822,000 with 99% probability, and the medium portfolio exceeds a wealth of $825,000 with 99% probability. The expected wealth of the cash-only portfolio is a mere $868,000 compared to the expected wealth of the medium portfolio of $1.538 million.

Table 5
Out-of-sample simulation results for CEW and improvement

	Exponential RA = 2		Increasing RRA		Decreasing RRA		Quadratic	
	CEW	Impr. (%)	CEW	Impr. (%)	CEW	Impr. (%)	CEW	Impr. (%)
Dynamic	1.41193		1.44004		1.35163		0.98204	
US stocks	1.28811	9.61	1.34365	7.17	0.68918	96.12	0.87639	12.06
Cash	0.86732	62.79	0.86728	66.04	0.86596	56.08	0.86628	13.36
Equally weighted	1.27082	11.10	1.28235	12.30	1.17981	14.56	0.96254	2.03
Averse	1.08667	29.93	1.08747	32.42	1.0605	27.45	0.97207	1.03
Medium	1.40415	0.55	1.4292	0.76	1.34326	0.62	0.97051	1.19
Prone	1.38933	1.63	1.43378	0.44	1.0925	23.72	0.93699	4.81

The risk-averse strategy exhibits the largest wealth of \$869,000 exceeded with 99% probability, with an expected wealth of \$1.098 million at the end of the investment horizon. In comparison, the dynamic strategy of decreasing relative risk aversion exhibits a very similar wealth of \$865,000 exceeded with 99% probability, but a significantly larger expected wealth of \$1.339 million. This represents an increase of 21.95% expected wealth at the same level of (downside) risk. The medium strategy exhibits the largest wealth of \$975,000 exceeded with 95% probability among all fixed-mix strategies. The best dynamic strategy on this measure is the quadratic downside risk strategy, which exceeds \$1.006 million with 95% probability. However, the quadratic strategy returns a significantly lower expected wealth of \$1.339 million, compared with \$1.538 million for the medium fixed-mix strategy. The extra downside protection is expensive on the upside. There is no dynamic strategy matching the 95% exceeded wealth of the medium fixed-mix strategy, and therefore we cannot make a statement of outperformance at the same level of 95% exceeded wealth given the strategies at hand.

In order to further evaluate the dynamic strategies, we compared the certainty equivalent wealth (CEW) of the dynamic strategies with the corresponding certainty equivalent wealth of each fixed-mix strategy. Table 5 presents the results in terms of certainty equivalent wealth and in the percentage improvement obtained when using the optimal dynamic strategy for each utility function compared to each of the fixed-mix strategies.

A certainty equivalent wealth (CEW) of \$1.41193 million is the maximum obtained from a dynamic strategy in the case of the exponential utility function with risk aversion of $\lambda = 2$. This utility represents a fairly aggressive investor. The medium strategy would be the best fixed-mix strategy with an CEW of \$1.40415 million. The improvement of the dynamic strategy versus the best fixed-mix strategy is 0.55%. The prone strategy would also fit reasonably well this investor, and results in a CEW of \$1.38933 million. The improvement of the dynamic strategy versus the prone strategy is 1.63%. The improvements in CEW of the dynamic strategy versus US stocks, cash and averse are 9.61, 62.79 and 29.93%, respectively.

Not surprisingly, the situation is very similar for the increasing relative risk aversion utility function. The dynamic strategy results in a certainty equivalent wealth of

$1.44004 million. The prone strategy is the best fixed-mix strategy with a CEW of $1.43378 million, resulting in an improvement of the dynamic strategy versus prone of 0.44%. The medium strategy is surpassed by the dynamic strategy by 0.76%. US stocks, cash only and risk averse are surpassed by the dynamic strategy by 7.17, 66.04 and 32.42%, respectively.

In the case of decreasing relative risk aversion, the dynamic strategy yields a CEW of $1.35163 million. The best fixed-mix strategy is the medium, with a CEW of $1.34326 million. The improvement of the dynamic strategy versus the medium strategy is 0.62%. The improvements of the dynamic strategy versus US stocks, cash-only, risk-averse and risk-prone are 95.12, 56.08, 27.45, and 23.72%, respectively, showing that none of these strategies are a particularly good fit for this type of utility function.

For the quadratic downside utility function, the dynamic strategy has a CEW of $0.98204 million, while the best fixed-mix performance is the risk-averse strategy with a CEW of $0.97207 million. Thus, the improvement of the dynamic strategy is 1.05% versus the best fixed-mix strategy. The improvements over US stocks, cash-only, medium and risk-prone are 12.06, 13.36, 1.19 and 4.81%, respectively.

The equally weighted strategy was never the best fixed-mix strategy for any utility function. For each utility function, the best fixed-mix strategy could be improved upon by about a half percent by the dynamic strategy, except for the quadratic utility function for which the improvement is about one percent. These results confirm that the improvement in certainty equivalent wealth can be significant. However, it also suggests that from a practical perspective a well-selected or numerically optimized fixed-mix strategy can be close in performance to a dynamic strategy when measured in terms of CEW. A possible explanation is that, when we used the historical mean return estimates, the difference in the mean returns between stock and corporate bonds was rather small, and the corresponding mean-variance efficient frontier rather flat. Using the mean estimates from the regression would result in an increased mean return for stocks, and consequently make the efficient frontier steeper. Thus, the benefits of dynamic asset allocation versus fixed-mix strategies would increase.

7.4. Dynamic strategies for hedging downside risk

In the next investment example, we consider an investment horizon of 10 years. An initial capital of $1 million is to be invested, and there are no further payments made in or out of the account. The capital should be invested so that a minimum return of 2% per year is exceeded with 99% probability, and the expected return should be as large as possible. Such an investment situation is often faced by institutional investors. Time horizons may vary. While the problem can be stated as a chance-constrained optimization problem, we will explore empirically whether downside risk protection can be achieved by utility maximization using the quadratic downside risk or the decreasing relative risk aversion utility function.

For the quadratic function we set the target wealth at $1.22 million reflecting the 2% per year minimum return to be exceeded with 99% probability. As in the investment

example above, we set the risk aversion at $\lambda_q/2 = 1000$, in order to heavily penalize any under-performance with respect to the target. For the decreasing relative risk aversion function, we set a high relative risk aversion of $\gamma = 8$ below the level $W_L = \$1.22$ million (the target wealth) and a low relative risk aversion of $\gamma = 1.01$ above the wealth level $W_H = \$1.5$ million, and assume the relative risk aversion increases linearly in between. For both utility functions, we solved the 10-year horizon problem assuming yearly, quarterly and monthly rebalancing. For quarterly rebalancing a 40-stage problem needed to be solved, and for the monthly rebalancing a 120-stage problem.

Table 6 gives the results for expected wealth, wealth exceeded with 99% probability, and minimum wealth (defined as the level exceeded with 99.9% probability), as well as expected return, return exceeded with 99% probability and minimum return (exceeded with 99.9% probability). Using yearly rebalancing, both utility functions lead to larger returns than 2% per annum with at least 99% probability. The quadratic downside utility function resulted in a very conservative strategy, and even the minimum returns (exceeded with 99.9% probability) met the 2% per year requirement. The expected returns from both strategies are 10.35% per year for the quadratic downside and 10.42% for the decreasing RRA utility function, indicating that the downside-risk-protected strategies are not overly costly on the upside. The results were rather similar for yearly, quarterly and monthly rebalancing for both utility functions; however, more frequent rebalancing allowed for somewhat better downside protection and larger expected returns.

For the quadratic downside risk strategy, Figure 13 represents the optimal asset allocation for yearly rebalancing with respect to wealth for one (top), five (middle) and nine years to go. For all attainable wealth levels (achieved with 99.9% probability), the optimal asset allocation is entirely on the right side of the critical wealth (with the most conservative asset allocation) and thus exhibits decreasing relative risk aversion. The better the performance (and the larger the wealth in the account), the more risk should be assumed and the larger should be the stock allocation. The attainable wealth in year 9 lies between \$1.16 and \$7.35 million, in year 5 between \$0.96 and \$3.54 million, and in

Table 6
Out-of-sample simulation results for downside risk protection

Rebal.	Periods	Exp. wealth	99% wealth	Min. wealth	Exp. ret. (%)	99% ret. (%)	Min. ret. (%)
Quadratic downside risk							
yearly	10	2.67628	1.31957	1.22778	10.35	2.81	2.07
quarterly	40	2.79864	1.28774	1.22500	10.84	2.56	2.05
monthly	120	2.71229	1.36413	1.26250	10.49	3.15	2.36
Decreasing RRA							
yearly	10	2.69504	1.25000	1.10000	10.42	2.26	0.96
quarterly	40	2.78378	1.26714	1.05000	10.78	2.40	0.49
monthly	120	2.76007	1.32083	1.10000	10.69	2.82	0.96

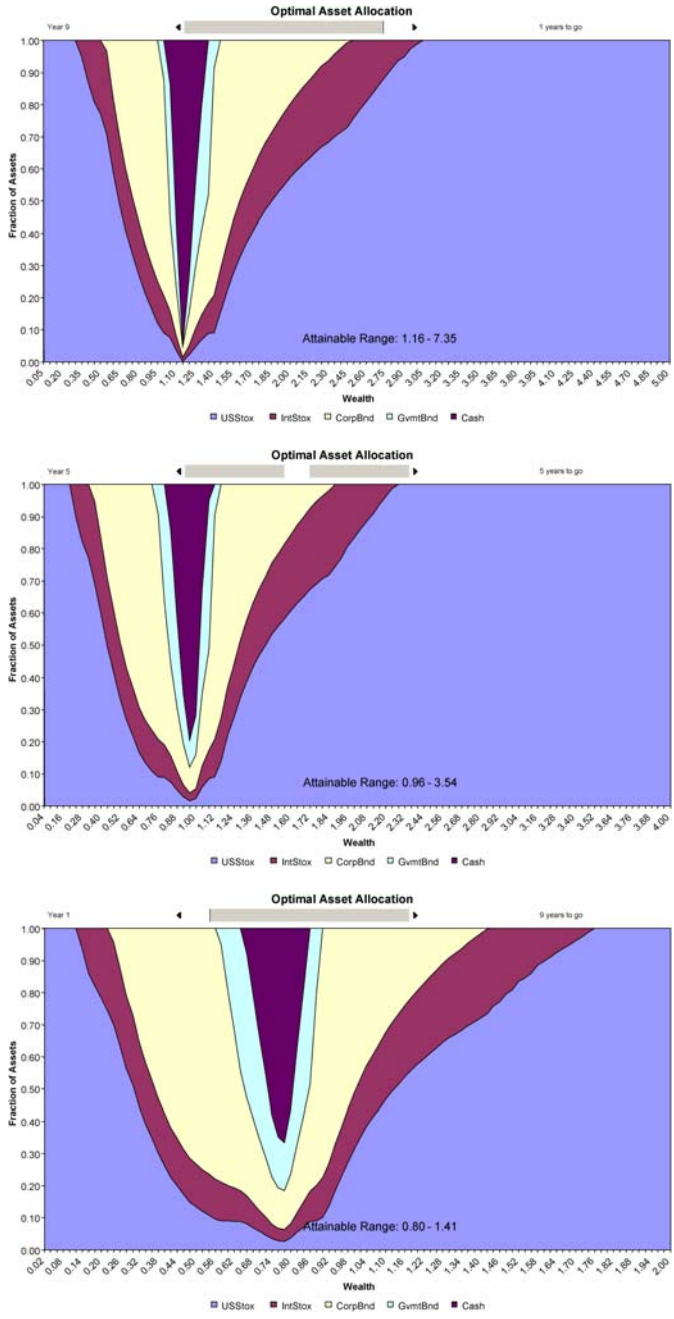


Fig. 13. Investment example: 10-year horizon: downside risk, quadratic, investment strategy.

Table 7
Out-of-sample simulation results for downside risk protection

Quadratic dynamic downside risk						
Period	Exp. wealth	99% wealth	Min. wealth	Exp. ret. (%)	99% ret. (%)	Min. ret. (%)
1	1.10008	0.8814	0.8	10.01	−11.86	−20.00
2	1.21161	0.89267	0.825	10.07	−5.52	−9.17
3	1.33354	0.92828	0.86	10.07	−2.45	−4.90
4	1.47004	0.97231	0.9	10.11	−0.70	−2.60
5	1.62245	1.01714	0.9616	10.16	0.34	−0.78
6	1.79528	1.07636	1	10.24	1.23	0.00
7	1.98402	1.1296	1.05333	10.28	1.76	0.74
8	2.19368	1.18286	1.12	10.32	2.12	1.43
9	2.42128	1.25091	1.16071	10.32	2.52	1.67
10	2.67628	1.31957	1.22778	10.35	2.81	2.07

year 1 between \$0.80 and \$1.41 million. Figure 14 (top) represents the expected value strategy (assuming expected returns are realized in every period), (middle) the simulations of wealth over time, and (bottom) the marginal distribution of wealth at the end of the investment horizon for yearly rebalancing. The expected value strategy prescribes to start with an initial asset allocation of 60% stocks and 40% corporate bonds, and then to gradually increase the stock allocation to almost 100% in the last rebalancing period. The marginal distribution is steep on the left side, representing a limited downside, and the minimum wealth (exceeded with 99.9%) of \$1.22778 million translates into a minimal annual return of 2.07%.

How well this strategy protects the downside along the way is presented in Table 7, which exhibits the expected wealth, the wealth exceeded with 99% probability, and the minimum wealth (that is exceeded with 99.9% probability), along with corresponding measures for the returns in each year of the investment horizon. The downside protection during the initial periods is rather weak, with the returns exceeded with 99% probability in year one resulting as −11.86% and the minimum return resulting as −20.00%. Only in year five is nominal capital preserved with greater than 99% probability and only in year six with 99.9% probability. In year eight the target return of 2% per year is exceeded with 99% probability, and in year 10 with 99.9% probability. The expected returns increase monotonically as time passes, but only marginally from 10.01% per year in year one to 10.35% in year ten.

We next compare these results to the results of a strategy where the downside is controlled dynamically in every period.

7.5. Downside risk protection at every period

Especially institutional investors are evaluated with regard to their portfolio's performance on a yearly basis. Given the ongoing evaluations, institutional managers may

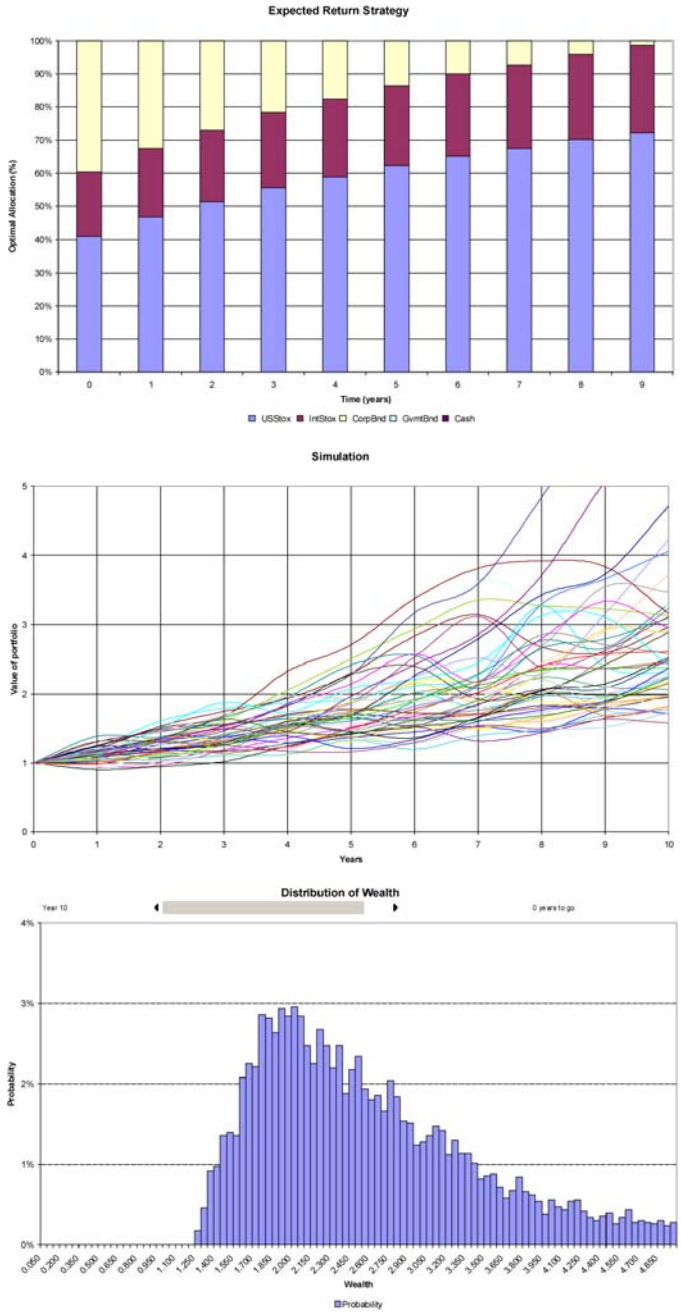


Fig. 14. Investment example: 10-year horizon: downside risk, quadratic, simulations.

consider implementing a dynamic strategy with limited downside risk not only at the end of the investment horizon but also at every year along the way. Again, using a 10-year investment horizon, we compute a strategy using a discounted additive quadratic downside utility function at every year. For the dynamic quadratic downside risk utility function, the discount factor was chosen as 1% per year and the target wealth in each period was set as $W_{d,t} = 1.02^t$, $t = 1{,}10$, reflecting the desired minimum return of 2% per year. We set the risk aversion coefficient at every period as $\lambda_l = 1000$ and $\lambda_q/2 = 1000$, in order to heavily penalize (linearly and quadratically) any under-performance with respect to the targets in addition to the quadratic downside utility function at the end of the investment horizon.

Figure 15 represents the strategies for one (top), five (middle) and nine (bottom) years to go. The optimal strategy at the last rebalancing period must be identical to that, where only terminal wealth is evaluated. This follows directly from the dynamic programming recursion. At all other years the dynamically controlled downside risk strategy is more conservative, e.g., at five years to go risk and stock allocation increases slower with wealth compared to the uncontrolled case (with downside risk protection only at the end of the investment horizon), and the most conservative allocation at the critical point is more conservative than in the uncontrolled case. For example, in the uncontrolled case at nine years to go, the cash position at the critical point is about 70%, whereas in the controlled case it is about 90%. With downside controls, the attainable wealth in year 9 lies between \$1.26 and \$5.36 million, in year 5 between \$1.13 and \$2.30 million, and in year 1 between \$1.01 and \$1.12 million. This is more narrow range than for the uncontrolled case, reflecting the loss on the upside induced by the protection on the downside.

Figure 16 (top) represents the expected value strategy (middle) the simulations of wealth, and (bottom) the marginal probability distribution of terminal wealth. The expected value strategy under downside control prescribes to start with a stock allocation of 5% and to increase it up to about 85% in year nine. This is significantly more conservative than in the uncontrolled case. Comparing the wealth simulations, one can readily observe that the downside is more protected along the way than in the uncontrolled case.

Table 8 presents the statistics for each year of the investment horizon. (The last line of year 10 represents the terminal statistics.) The marginal distribution of final wealth displays a minimum wealth (exceeded with 99.9% probability) of \$1.3125 million, which translates into a 2.76% annual return. Compared to the uncontrolled case above, the downside is much better protected along the way, e.g., in every period the return exceeded with 99% probability is greater than 2%, and the minimum return (exceeded with 99.9% probability) exceeds 2% from year three on. The expected return is smaller compared to the unconstrained case and increases monotonically from 6.58 to 9.22%, which confirms that the dynamically downside-controlled strategy is overall more conservative than when only the terminal distribution is considered in the optimization.

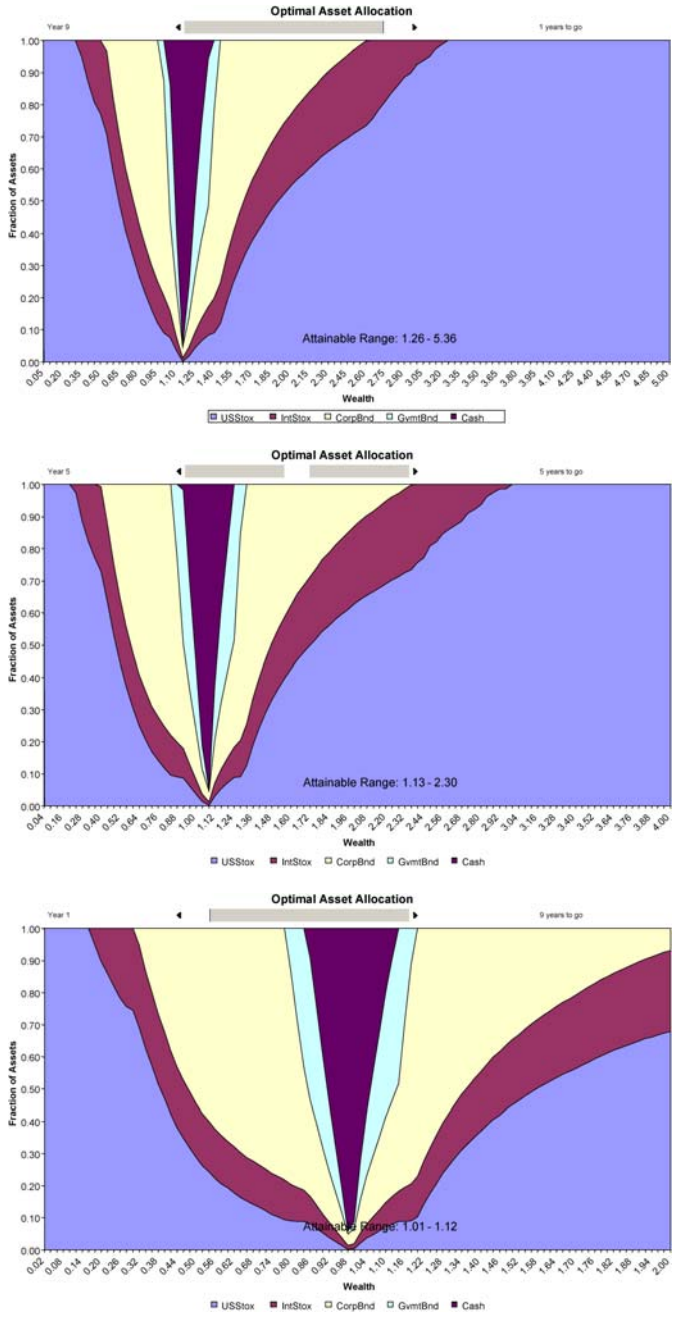


Fig. 15. Investment example: 10-year horizon: downside risk at each period, quadratic, investment strategy.

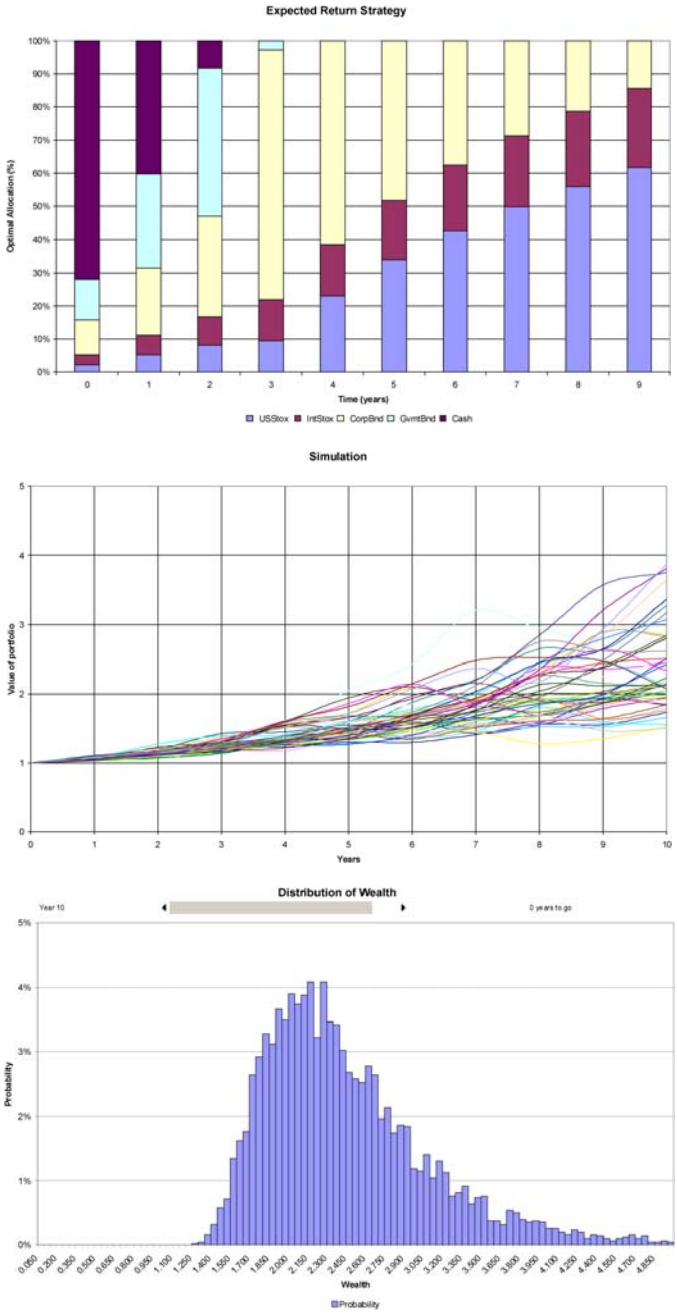


Fig. 16. Investment example: 10-year horizon: downside risk at each period, quadratic, simulations.

Table 8
Out-of-sample simulation results for dynamic downside risk protection at every period

Quadratic dynamic downside risk						
Period	Exp. wealth	99% wealth	Min. wealth	Exp. ret. (%)	99% ret. (%)	Min. ret. (%)
1	1.06575	1.02225	1.00533	6.58	2.23	0.53
2	1.14561	1.06108	1.032	7.03	3.01	1.59
3	1.24271	1.09137	1.06222	7.51	2.96	2.03
4	1.35621	1.13641	1.095	7.91	3.25	2.29
5	1.48549	1.176	1.12667	8.24	3.30	2.41
6	1.63401	1.22638	1.152	8.53	3.46	2.39
7	1.79975	1.27385	1.2	8.76	3.52	2.64
8	1.9845	1.32121	1.23	8.94	3.54	2.62
9	2.18661	1.37778	1.26	9.08	3.62	2.60
10	2.41527	1.43966	1.3125	9.22	3.71	2.76

These are very encouraging results. All out-of-sample simulations were performed using bootstrapping from the historical monthly observations and thus represent possible fat tails in the asset returns distribution.

7.6. Computation times

For the optimization we used a sample size of $|S_t^i| = 2000$ of annual returns for each period. The sample is based on bootstrapping from the monthly returns, thus we used 24,000 sampled monthly observations to compute the annual returns sample. In each period for the out-of-sample evaluation, we used a sample size of $|S_t^o| = 5000$ of annual returns, also obtained by bootstrapping monthly returns. The computations were carried out using the fast method, pre-calculating the efficient frontier parameterized at 2000 points. After computing the optimal strategy, we simulated 5000 paths based on bootstrapping monthly returns. Table 9 presents the optimization time in CPU seconds (on a Pentium 4 3.4 MHz and 2 GB of RAM) for the different time horizons and utility functions used. It takes about 11 seconds to optimize a 20-period problem with yearly rebalancing, thus solving a 20-stage stochastic dynamic program. The computation time was about 8 seconds for a 10-year horizon with yearly rebalancing, about 17 seconds for quarterly rebalancing and 41 seconds for monthly rebalancing. In the latter, a 120 stage stochastic dynamic program was solved in 41 seconds.

The fast computations for yearly rebalancing permit us to carry out the calculations on-line. We have developed a commercial system, called WealthiOR, to help investment advisors find the best asset allocation strategies for their clients, individual and institutional. The system is web-based and the optimizations and simulations are carried out on-line in real time. The system is user friendly, easy to use, and also facilitates graphical output and downloadable tables for spreadsheets, providing all necessary sta-

Table 9
Computation times in elapsed CPU seconds

Rebalancing	Horizon	Periods	Utility	
			Exponential	Increasing RRA
yearly	20	20	10.53	10.92
			Quadratic	Decreasing RRA
yearly	20	20	10.48	11.81
			Quadratic	Decreasing RRA
yearly	10	10	6.58	7.79
quarterly	10	40	15.85	17.06
monthly	10	120	41.35	41.35

tistical information. It permits an adviser to quickly devise customized (regarding risk preference and time horizon) optimized forward-looking dynamic strategies.

8. Comparison to multi-stage stochastic programming

Our results demonstrate that, using the stochastic dynamic programming approach, asset allocation problems with many asset classes and many time periods can be solved efficiently and accurately, allowing for a detailed discussion of the resulting investment strategies. The approach can also handle options and serial dependency of asset returns modeled by restricted autoregressive processes using a small number of predictors. Also, stochastic cash flows to represent uncertain liabilities can be handled by the approach to the extent that the serial correlation of the joint probability distribution of asset returns and cash flows can be modeled with a small number of state variables. This makes the approach also applicable to classes of the more general asset-liability problem. From a statistical perspective (considering the historical data available for estimating the processes and the number of parameters that need to be estimated), predictions based on too many autoregressive factors may lead to over-fitting, resulting in weak out-of-sample performance. Thus, from a practical perspective, predictions based on a limited number of factors (state variables) seem adequate, and thus the stochastic dynamic programming approach can be efficiently employed for solving a wide class of practical problems.

For general returns processes with many predictive variables, and when transaction costs must be considered, multi-stage stochastic programming can be employed successfully. For example, Collomb and Infanger (2005) explored the impact of serial correlation in asset returns on the optimal asset allocation strategy. The focus of the study was on active portfolio management on the asset class level in a multi-period look-ahead setting. For representing serial correlation, we estimated a lag-one vector autoregressive process for stocks, bonds and cash, and formulated a four-stage stochastic linear program representing the dynamic asset allocation problem for three annual

periods of look-ahead. Modeling the stochastic program, a typical scenario tree was represented by (50 × 50 × 40) scenarios leading to 100,000 scenarios, and the corresponding deterministic equivalent formulation of the stochastic program had several hundred thousand rows and columns. Such problems were typically solved in about 2 hours CPU time as deterministic equivalent linear programs and in about 20 minutes when solved using decomposition as implemented in Infanger's (1997) DECIS code. In comparison, the stochastic dynamic programming recursion with 2000 scenarios per period and 10 periods look ahead took only about 10 seconds.

Acknowledgements

The author wishes to thank Peter Glynn, John Stone, and the editors, Stavros A. Zenios and William T. Ziemba, for their editorial suggestions on previous versions of this chapter.

References

Adaci, J., 1996. Combining stochastic linear programming and dynamic programming to solve multiperiod portfolio problems, Dissertation. Department of Operations Research, Stanford University.

Arrow, K.J., 1971. Essays on the Theory of Risk Bearing. Markham, Chicago.

Bell, D., 1988. One switch utility functions and a measure of risk. Management Science 24 (12), 14126–14240.

Bellman, R., 1957. Dynamic Programming. Princeton University Press, Princeton, NJ.

Bernoulli, D., 1738. Specimen theoriae novae de mensura sortis. In: Commentarii Academiae Scientiarum Imperiales Petropolitanae, vol. 5, pp. 175–192. Translated by L. Sommer (1954). Exposition of a new theory on the measurement of risk. Econometrica 22, 22–36.

Blake, D., Lehmann, B.N., Timmermann, A., 1999. Asset allocation dynamics and pension fund performance. Journal of Business 72 (4), 429–461.

Bodie, Z., Crane, D.B., 1997. Personal investing: Advice, theory, and evidence. Financial Analyst Journal 53 (6), 13–23.

Brennan, M.J., Schwartz, E.S., 1998. The use of treasury bill futures in strategic asset allocation programs. In: Ziemba, W.T., Mulvey, J.M. (Eds.), World Wide Asset and Liability Modeling. Cambridge University Press, Cambridge, pp. 205–228.

Brennan, M.J., Schwartz, E.S., Lagnado, R., 1997. Strategic asset allocation. Journal of Economic Dynamics and Control 21, 1377–1403.

Brooke, A., Kendrik, D., Meeraus, A., 1988. GAMS, A Users Guide. The Scientific Press, South San Francisco, CA.

Campbell, J.Y., Viceira, L.M., 2002. Strategic Asset Allocation. Oxford University Press, Oxford, NY.

Carino, D.R., Kent, T., Myers, D.H., Stacy, C., Sylvanus, M., Turner, A.L., Watanabe, K., Ziemba, W.T., 1994. The Russel–Yasuda Kasai model: An asset-liability model for a Japanese insurance company using multistage stochastic programming. Interfaces 24 (1), 29–49.

Chacko, G., Neumar, K., 2006. Perturbation methods for dynamic portfolio allocation problems. In: Zenios, S.A., Ziemba, W.T. (Eds.), In: Handbook of Asset and Liability Management: Theory and Methodology, vol. 1. Elsevier, Amsterdam, pp. 337–384.

Chopra, V.K., 1993. Improving optimization. Journal of Investing 2 (3), 51–59.

Chopra, V.K., Ziemba, W.T., 1993. The effect of errors in mean, variances and covariances on optimal portfolio choice. Journal of Portfolio Management, 6–11, Winter.

Collomb, A., Infanger, G., 2005. The impact of serial correlation of asset returns on dynamic asset allocation strategies, SOL Technical Report.

Connor, G., 1997. Sensible return forecasting for portfolio management. Financial Analysts Journal (September/October), 44–51.

Consigli, G., 2006. Asset liability management for individual investors. In: Zenios, S.A., Ziemba, W.T. (Eds.), In: Handbook of Asset and Liability Management: Applications and Case Studies, vol. 2. Elsevier, Amsterdam.

Consiglio, A., Cocco, F., Zenios, S.A., 2006. The Prometaia Calcolo model for endowments with guarantees. In: Zenios, S.A., Ziemba, W.T. (Eds.), In: Handbook of Asset and Liability Management: Applications and Case Studies, vol. 2. Elsevier, Amsterdam.

Constantinides, G.M., 2002. Rational asset prices. The Journal of Finance 57 (4), 1567–1591.

Cox, J.C., Huang, C., 1999. Optimal consumption and portfolio policies when asset prices follow a diffusion process. Journal of Economic Theory 49 (October), 33–83.

Dantzig, G.B., Glynn, P.W., 1990. Parallel processors for planning under uncertainty. Annals of Operations Research 22, 1–21.

Dantzig, G.B., Infanger, G., 1993. Multi-stage stochastic linear programs for portfolio optimization. Annals of Operations Research 45, 59–76.

Dembo, R., Mausser, H., 2000. The put/call efficient frontier. Algo Research Quarterly 3 (1), 13–25.

De Farias, D.P., Van Roy, B., 2003. The linear programming approach to approximate dynamic programming. Operations Research 51 (6), 850–865.

Dimson, E., Marsh, P., Staunton, M., 2002. Triumph of the Optimists: 101 Years of Global Investment Returns. Princeton University Press, Princeton, NJ.

Edirishinghe, N.C.P., 2006. Integrated risk control using stochastic programming. In: Zenios, S.A., Ziemba, W.T. (Eds.), In: Handbook of Asset and Liability Management: Applications and Case Studies, vol. 2. Elsevier, Amsterdam.

Faig, M., 2002. How do households invest their wealth. Report. Centre de Recerca en Economia Interactional, Generalitat de Catalunya, Universitat Pompeu Fabra.

Frauendorfer, K., 1996. Barycentric scenario trees in convex multistage stochastic programming. Mathematical Programming 75, 277–293.

Friend, I., Blume, M., 1975. The demand for risky assets. American Economic Review 65 (5), 900–922.

Golub, B., Holmer, M., McKendall, R., Pohlman, L., Zenios, S.A., 1995. A stochastic programming model for money management. European Journal of Operational Research 85, 282–296.

Grinold, R.C., 1999. Mean variance and scenario based approaches to portfolio selection. Journal of Portfolio Management, 10–22, Winter.

Gunthorpe, D., Levy, H., 1994. Portfolio composition and the investment horizon. Financial Analyst Journal (January–February), 51–56.

Hakansson, N.H., 1971. On myopic portfolio policies, with and without serial correlation of yields. Journal of Business 44 (3), 324–334.

Hanoch, G., Levy, H., 1969. The efficiency analysis of choices involving risk. Review of Economic Studies 36 (107), 335–346.

Hiller, R.S., Eckstein, J., 1993. Stochastic dedication: Designing fixed income portfolios using massively parallel Benders decomposition. Management Science 39 (11), 1422–1438.

Infanger, G., 1992. Monte Carlo (importance) sampling within a Benders decomposition algorithm for stochastic linear programs. Annals of Operations Research 39, 41–67.

Infanger, G., 1994. Planning Under Uncertainty—Solving Large-Scale Stochastic Linear Programs. The Scientific Press Series. Boyd and Fraser.

Infanger, G., 1997. DECIS User's Guide, Dr. Gerd Infanger, 814 Loma Verde Ave, Palo Alto, CA 94303.

Infanger, G., 1999. Managing risk using multi-stage stochastic optimization, Report SOL 99-2, Dept. of Operations Research, Stanford University, CA and The Institute for Quantitative Research in Finance

("The Q-Group"), INQUIRE-United Kingdom, and INQUIRE-Europe, Proceedings of the Autumn 2000 Joint Conference.

Infanger, G., Morton, D., 1996. Cut sharing for multistage stochastic linear programs with interstage dependency. Mathematical Programming 75, 241–256.

Kahnemann, D., Tversky, A., 1979. Prospect theory: An analysis of decisions under risk. Econometrica 47 (2), 263–291.

Kallberg, J.G., Ziemba, W.T., 1981. Remarks on optimal portfolio selection. In: Bamberg, G., Opitz, O. (Eds.), Methods of Operations Research, vol. 44. Oelgeschlager, Gunn and Hain, pp. 507–520.

Kallberg, J.G., Ziemba, W.T., 1983. Comparison of alternative utility functions in portfolio selection problems. Management Science 29, 1257–1276.

Kallberg, J.G., Ziemba, W.T., 1984. Mis-specification in portfolio selection problems. In: Bamberg, G., Spreemann, K. (Eds.), Risk and Capital. Springer-Verlag, Berlin/New York, pp. 74–87.

Keeney, R., Raiffa, H., 1976. Decisions with Multiple Objectives. John Wiley and Sons, New York.

Kouwenberg, R., Zenios, S.A., 2006. Stochastic programming models for asset liability management. In: Zenios, S.A., Ziemba, W.T. (Eds.), Handbook of Asset and Liability Management: Theory and Methodology, vol. 1. Elsevier, Amsterdam, pp. 253–303.

Kusy, M.I., Ziemba, W.T., 1986. A bank asset and liability management model. Operations Research 35, 356–376.

Levy, H., 1972. Portfolio performance and the investment horizon. Management Science 18 (12), B-645–B-653.

Markowitz, H.M., 1952. Portfolio selection. The Journal of Finance 7, 77–91.

Markowitz, H.M., 1994. The value of a blank check. Journal of Portfolio Management, 82–91, Summer.

Markowitz, H.M., van Dijk, E., 2006. Risk-return analysis. In: Zenios, S.A., Ziemba, W.T. (Eds.), Handbook of Asset and Liability Management: Theory and Methodology, vol. 1. Elsevier, Amsterdam, pp. 139–197.

MacLean, L., Zhao, Y., Ziemba, W.T., 2006. Dynamic portfolio selection with process control. Journal of Banking and Finance 30, 317–339. Special issue edited by P. Krokhmal, R.T. Rockafellar, S. Uryasev.

MacLean, L., Ziemba, W.T., Li, Y., 2005. Time to wealth goals in capital accumulation, Quantitative Finance, submitted for publication.

Merton, R.C., 1969. Lifetime portfolio selection under uncertainty: the continuous-time case. Review of Economics and Statistics 51 (3), 247–259.

Merton, R.C., 1990. Continuous-Time Finance. Blackwell Publishing, 350 Main Street, Malden, MA 02148-5020, USA.

Michaud, R.O., 1989. The Markowitz optimization enigma: Is "optimized" optimal? Financial Analysts Journal (January/February), 31–42.

Mossin, J., 1968. Optimal multiperiod portfolio policies. Journal of Business, 215–229.

Mulvey, J.M., Vladimirou, H., 1992. Stochastic network programming for financial planning problems. Management Science 38, 1642–1664.

Mulvey, J.M., Pauling, B., Britt, S., Morin, F., 2006. Dynamic financial analysis for multinational insurance companies. In: Zenios, S.A., Ziemba, W.T. (Eds.), In: Handbook of Asset and Liability Management: Applications and Case Studies, vol. 2. Elsevier, Amsterdam.

Musumeci, J., Musumeci, J., 1999. A dynamic programming approach to multiperiod asset allocation. Journal of Financial Services Research 15 (1), 5–21.

Nielsen, S.S., Zenios, S.A., 1996. A stochastic programming model for funding single-premium deferred annuities. Mathematical Programming 75, 177–200.

Pratt, J.W., 1964. Risk aversion in the small and in the large. Econometrica 32, 122–136.

Rubinstein, M., 1976. A strong case for generalized logarithmic utility model as the premier model of financial markets. The Journal of Finance 31 (2), 551–571.

Samuelson, P., 1969. Lifetime portfolio selection by dynamic stochastic programming. Review of Economics and Statistics 51, 239–246.

Siegel, J., 2002. Stocks for the Long Run: The Definitive Guide to Financial Market Returns and Long-Term Investment Strategies. McGraw-Hill Trade, New York.

Tobin, J., 1958. Liquidity preference as behavior towards risk. Review of Economic Studies 25 (2), 65–86.

Treynor, J.L., 2003. Time diversification. Journal of Investment Management 2 (3), 36–47.

Von Neumann, J., Morgenstern, O., 1944. Theory of Games and Economic Behavior. Princeton University Press, Princeton, NJ.

Wallace, S.W., Ziemba, W.T. (Eds.), 2005. Applications of Stochastic Programming. SIAM – Mathematical Programming Society, Philadelphia, PA.

Zenios, S.A., 1993. A model for portfolio management with mortgage-backed securities. Annals of Operations Research 43, 337–356.

Ziemba, W.T., 1974. Choosing investment portfolios when the returns have stable distributions. In: Hammer, P.L., Zoutendijk, G. (Eds.), Mathematical Programming in Theory and Praxis. North-Holland, Amsterdam, pp. 443–482.

Ziemba, W.T., 2003. The Stochastic Programming Approach to Asset, Liability, and Wealth Management. The Research Foundation of AIMR, Charlottesville, VA.

Ziemba, W.T., 2006. The Russell Yasuda, InnoALM, and related models for pensions, insurance companies and high net worth individuals. In: Zenios, S.A., Ziemba, W.T. (Eds.), In: Handbook of Asset and Liability Management: Applications and Case Studies, vol. 2. Elsevier, Amsterdam.

Ziemba, W.T., Parkan, C., Brooks-Hill, R., 1974. Calculation of investment portfolios with risk free borrowing and lending. Management Science 21 (2), 209–222.

This page intentionally left blank

Chapter 6

STOCHASTIC PROGRAMMING MODELS FOR ASSET LIABILITY MANAGEMENT

ROY KOUWENBERG[*]

Asian Institute of Technology, School of Management, P.O. Box 4, Klong Luang, Pathumthani 12120, Thailand
e-mail: kouwenbe@ait.ac.th

STAVROS A. ZENIOS[†]

HERMES Center on Computational Finance & Economics, University of Cyprus, P.O. Box 20537, 1678 Nicosia, Cyprus
and
Financial Institutions Center, The Wharton School, Philadelphia, USA
e-mail: zenioss@ucy.ac.cy

Contents

Abstract 254
1. Introduction 255
2. Stochastic programming 256
   2.1. Basic concepts in stochastic programming 256
      2.1.1. Anticipative models 256
      2.1.2. Adaptive models 257
      2.1.3. Recourse models 258
      2.1.4. Deterministic equivalent formulation 259
      2.1.5. Multistage models 260
   2.2. Stochastic programming model for portfolio management 261
      2.2.1. Notation 262
      2.2.2. Model formulation 264
3. Scenario generation and tree construction 267
   3.1. Scenarios for the liabilities 267
   3.2. Scenarios for economic factors and asset returns 270
   3.3. Methods for generating scenarios 272
      3.3.1. Bootstrapping historical data 272

[*] This chapter was written while Roy Kouwenberg was visiting the University of British Columbia, Faculty of Commerce. Research partially supported by NSERC grant A7147.
[†] Research partially supported by EC grant ICA1-CT-2000-70015.

Handbook of Asset and Liability Management, Volume 1
Edited by S.A. Zenios and W.T. Ziemba
Copyright © 2006 Elsevier B.V. All rights reserved
DOI: 10.1016/S1872-0978(06)01006-4

3.3.2. Statistical models from the Value-at-Risk literature 272
3.3.3. Modelling economic factors and asset returns for a pension fund 274
3.4. Constructing event trees 277
3.4.1. Random sampling 277
3.4.2. Adjusted random sampling 279
3.4.3. Fitting the mean and the covariance matrix 280
3.5. Options, bonds and arbitrage 283
3.5.1. Arbitrage-free event trees 284
4. Comparison of stochastic programming with other methods 287
4.1. Mean-variance models and downside risk 287
4.2. Discrete-time multi-period models 288
4.3. Continuous-time models 290
4.4. Stochastic programming 291
5. Applications of stochastic programming to ALM 291
6. Solution methods and computations 296
7. Summary and open issues 297
Appendix A. Basics of probability spaces 298
References 299

Abstract

Stochastic programming is a powerful modelling paradigm for asset and liability management problems. It incorporates in a common framework multiple correlated sources of risk for both the asset and liability side, takes a long time horizon perspective, accommodates different levels of risk aversion and allows for dynamic portfolio rebalancing while satisfying operational or regulatory restrictions and policy requirements. This chapter introduces stochastic programming models for broad classes of asset and liability management problems, describes procedures for generating the requisite event trees, discusses the validity of model results for illustrative applications, compares stochastic programming with alternative modelling approaches, and hinges upon solution techniques and computational issues.

1. Introduction

Asset and liability management (abbreviated: ALM) problems deal with uncertainty. They deal with the planning of financial resources in the face of uncertainty about economic, capital market, actuarial, and demographic conditions. A general approach for dealing with uncertain data is to assign to the unknown parameters a probability distribution, which should then be incorporated into an appropriate mathematical programming model. Mathematical programming models for dealing with uncertainty are known as *stochastic programs*. Stochastic programming is recognized as a powerful modelling paradigm for several areas of application. Its validity for ALM problems, in particular, is enhanced by that fact that it readily incorporates in a common framework multiple correlated sources of risk for both the asset and liability side, has long time horizons, accommodates risk aversion, and allows for dynamic portfolio rebalancing while satisfying operational or regulatory restrictions and policy requirements. Thus it facilitates an integrated view of the risk management process at an enterprise-wide level (Holmer and Zenios, 1995; Chapter 1 of this handbook). The applicability of stochastic programming for financial planning was recognized first by Bradley and Crane (1972) and Ziemba and Vickson (1975). But it was not until the nineties that stochastic programming started gaining prominence as a decision support tools for asset and liability management. This development was motivated in part by algorithmic advances that enabled the solution of large-scale realistic models. Globalization and innovations in the financial markets are the driving force behind the development of stochastic programming models for ALM that continues unabated to this date, aided by advances in computing technology and the availability of software.

Several academic researchers and practitioners demonstrated the effectiveness of stochastic programming models in supporting financial decision making. We mention the most recent contributions: Kusy and Ziemba (1986) for bank management, Mulvey and Vladimirou (1992) for asset allocation, Zenios (1991, 1995), Golub et al. (1995) and Nielsen and Zenios (1996a, 1996b) for fixed income portfolio management, Carino et al. (1994), Carino and Ziemba (1998), Consigli and Dempster (1998), Høyland (1998) and Mulvey, Gould and Morgan (2000) for insurance companies, Dert (1995) for pension funds, Consiglio, Cocco and Zenios (2001) for minimum guarantee products. These are some of the applications that were done jointly with commercial sponsors and were adopted in practical settings. See also the case studies in this handbook; additional references are given in Zenios (1993, 2006) and Ziemba and Mulvey (1998).

This chapter reviews stochastic programming models for asset and liability management. Section 2 introduces the basics of stochastic programming and formulates a canonical model for portfolio management. The key issue of generating probabilistic data for a stochastic programming ALM system is elaborated in Section 3 on scenario generation methods. The performance of a stochastic programming ALM model for pension funds is also discussed in this section, in conjunction with alternative scenario generation methods. Section 4 places stochastic programming models in the context of the traditional portfolio choice literature from financial economics, and discusses its

advantages and limitations. A brief review of stochastic programming applications to ALM in several institutional settings is given in Section 5, and references are also made to models from other chapters of this handbook. Section 6 hinges upon solution techniques illustrating the size of problems that are solvable with current state-of-the-art software. Open issues are discussed in Section 7.

2. Stochastic programming

Stochastic programming models were first formulated as mathematical programs in the mid 1950s, independently, by Dantzig (1955) and Beale (1955). Modern textbook treatments of stochastic programming are Kall and Wallace (1994) and Birge and Louveaux (1997), and research literature is given in the chapter by Wets (1989) or the book by Censor and Zenios (1997) which focuses on solution methods. We introduce the basics of stochastic programming and then formulate a canonical model for portfolio management. The appendix gives some background on probability theory which is essential in understanding stochastic programming with continuous random variables. Readers interested in the resulting large-scale nonlinear programs defined using discrete and finite scenario sets can do without this background.

2.1. Basic concepts in stochastic programming

We formulate first two special cases of stochastic programs, the *anticipative* and the *adaptive* models. We then combine the two in the most general formulation of the *recourse* model which is the one suited for financial applications. Boldface Greek characters are used to denote random vectors which belong to some probability space as defined in Appendix A.

2.1.1. Anticipative models

Consider the situation where a decision x must be made in an uncertain world where the uncertainty is described by the random vector $\boldsymbol{\omega}$. The decision does not in any way depend on future observations, but prudent planning has to anticipate possible future realizations of the random vector.

In anticipative models feasibility is expressed in terms of *probabilistic* (or *chance*) *constraints*. For example, a *reliability* level α, where $0 < \alpha \leqslant 1$, is specified and constraints are expressed in the form

$$P\big\{\boldsymbol{\omega} \mid f_j(x, \boldsymbol{\omega}) = 0,\ j = 1, 2, \ldots, n\big\} \geqslant \alpha,$$

where x is the m-dimensional vector of decision variables and $f_j : \mathbb{R}^m \times \Omega \to \mathbb{R}$, $j = 1, 2, \ldots, n$. The objective function may also be of a reliability type, such as $P\{\boldsymbol{\omega} \mid f_0(x, \boldsymbol{\omega}) \leqslant \gamma\}$, where $f_0 : \mathbb{R}^m \times \Omega \to \mathbb{R} \cup \{+\infty\}$ and γ is a constant.

An anticipative model selects a policy that leads to some desirable characteristics of the constraint and objective functionals under the realizations of the random vector. In the example above it is desirable that the probability of a constraint violation is less than the prespecified threshold value α. The precise value of α depends on the application at hand, the cost of constraint violation, and other similar considerations.

2.1.2. Adaptive models

In an adaptive model observations related to uncertainty become available before a decision x is made, such that optimization takes place in a learning environment. It is understood that observations provide only partial information about the random variables because otherwise the model would simply wait to observe the values of the random variables, and then make a decision x by solving a deterministic mathematical program. In contrast to this situation we have the other extreme where all observations are made after the decision x has been made, and the model becomes anticipative.

Let $\mathcal{A}$ be the collection of all the relevant information that could become available by making an observation. This $\mathcal{A}$ is a subfield of the σ-field (see Appendix A) of all possible events, generated from the support set Ω of the random vector $\boldsymbol{\omega}$. The decisions x depend on the events that could be observed, and x is termed $\mathcal{A}$-*adapted* or $\mathcal{A}$-*measurable*. Using the conditional expectation with respect to $\mathcal{A}$, $E[\cdot|\mathcal{A}]$, the adaptive stochastic program can be written as:

$$\begin{aligned} &\text{Minimize} && E\big[f_0(x(\boldsymbol{\omega}),\boldsymbol{\omega})|\mathcal{A}\big] \\ &\text{subject to} && E\big[f_j(x(\boldsymbol{\omega}),\boldsymbol{\omega})|\mathcal{A}\big] = 0, \quad j = 1,2,\ldots,n, \\ &&& x(\boldsymbol{\omega}) \in X, \quad \text{almost surely.} \end{aligned} \tag{1}$$

The mapping $x : \Omega \to X$ is such that $x(\boldsymbol{\omega})$ is $\mathcal{A}$-measurable. This problem can be addressed by solving for every $\boldsymbol{\omega}$ the following deterministic programs:

$$\text{Minimize} \quad E\big[f_0(x,\cdot)|\mathcal{A}\big](\boldsymbol{\omega}) \tag{2}$$

$$\text{subject to} \quad E\big[f_j(x,\cdot)|\mathcal{A}\big](\boldsymbol{\omega}) = 0, \quad j = 1,2,\ldots,n, \tag{3}$$

$$x \in X. \tag{4}$$

The two extreme cases (i.e., complete information with $\mathcal{A} = \Sigma$, or no information at all) deserve special mention. The case of no information reduces the model to the form of the anticipative model; when there is complete information model (1) is known as the *distribution model*. The goal in this later case is to characterize the distribution of the optimal objective function value. The precise values of the objective function and the optimal policy x are determined after realizations of the random vector $\boldsymbol{\omega}$ are observed. The most interesting situations arise when partial information becomes available after some decisions have been made, and models to address such situations are discussed next.

2.1.3. Recourse models

The *recourse problem* combines the anticipative and adaptive models in a common mathematical framework. The problem seeks a policy that not only anticipates future observations but also takes into account that observations are made about uncertainty, and thus can adapt by taking *recourse* decisions. For example, a portfolio manager specifies the composition of a portfolio considering both future movements of stock prices (anticipation) and that the portfolio will be rebalanced as prices change (adaptation).

The two-stage version of this model is amenable to formulations as a large-scale deterministic nonlinear program with a special structure of the constraint matrix. To formulate the two-stage stochastic program with recourse we need two vectors for decision variables to distinguish between the anticipative policy and the adaptive policy. The following notation is used.

$x \in \mathbb{R}^{m_0}$ denotes the vector of *first-stage* decisions. These decisions are made before the random variables are observed and are anticipative.

$y \in \mathbb{R}^{m_1}$ denotes the vector of *second-stage* decisions. These decisions are made after the random variables have been observed and are adaptive. They are constrained by decisions made at the first-stage, and depend on the realization of the random vector.

We formulate the *second-stage problem* in the following manner. Once a first-stage decision x has been made, some realization of the random vector can be observed. Let $q(y, \boldsymbol{\omega})$ denote the second-stage cost function, and let $\{T(\boldsymbol{\omega}), W(\boldsymbol{\omega}), h(\boldsymbol{\omega}) \mid \boldsymbol{\omega} \in \Omega\}$ be the model parameters. Those parameters are functions of the random vector $\boldsymbol{\omega}$ and are, therefore, random parameters. T is the *technology matrix* of dimension $n_1 \times m_0$. It contains the technology coefficients that convert the first-stage decision x into resources for the second-stage problem. W is the *recourse matrix* of dimension $n_1 \times m_1$. h is the second-stage resource vector of dimension n_1.

The second-stage problem seeks a policy y that optimizes the cost of the second-stage decision for a given value of the first-stage decision x. We denote the optimal value of the second-stage problem by $\mathcal{Q}(x, \boldsymbol{\omega})$. This value depends on the random parameters and on the value of the first-stage variables x. $\mathcal{Q}(x, \boldsymbol{\omega})$ is the optimal value, for any given Ω, of the following nonlinear program

$$\begin{aligned} &\text{Minimize} && q(y, \boldsymbol{\omega}) \\ &\text{subject to} && W(\boldsymbol{\omega})y = h(\boldsymbol{\omega}) - T(\boldsymbol{\omega})x, \\ &&& y \in \mathbb{R}^{m_1}_+. \end{aligned} \tag{5}$$

If this second-stage problem is infeasible then we set $\mathcal{Q}(x, \boldsymbol{\omega}) \doteq +\infty$. The model (5) is an *adaptation* model in which y is the *recourse* decision and $\mathcal{Q}(x, \boldsymbol{\omega})$ is the *recourse cost function*.

The two-stage stochastic program with recourse is an optimization problem in the first-stage variables x, which optimizes the sum of the cost of the first-stage decisions,

$f(x)$, and the expected cost of the second-stage decisions. It is written as follows:

$$\begin{aligned} &\text{Minimize} && f(x) + E\big[\mathcal{Q}(x, \boldsymbol{\omega})\big] \\ &\text{subject to} && Ax = b, \\ & && x \in \mathbb{R}^{m_0}_+, \end{aligned} \tag{6}$$

where A is an $n_0 \times m_0$ matrix of constraint coefficients, and b is an n_0-vector denoting available resources at the first stage.

Combining (5) and (6) we obtain the following model:

$$\begin{aligned} &\text{Minimize} && f(x) + E\Big[\operatorname*{Min}_{y \in \mathbb{R}^{m_1}_+} \big\{q(y, \boldsymbol{\omega}) \mid T(\boldsymbol{\omega})x + W(\boldsymbol{\omega})y = h(\boldsymbol{\omega})\big\}\Big] \\ &\text{subject to} && Ax = b, \\ & && x \in \mathbb{R}^{m_0}_+. \end{aligned} \tag{7}$$

("Min" denotes the minimal function value.)

Let $K_1 \doteq \{x \in \mathbb{R}^{m_0}_+ \mid Ax = b\}$, denote the feasible set for the first-stage problem. Let also $K_2 \doteq \{x \in \mathbb{R}^{m_0} \mid E[\mathcal{Q}(x, \boldsymbol{\omega})] < +\infty\}$ denote the set of *induced constraints*. This is the set of first-stage decisions x for which the second-stage problem is feasible. Problem (6) is said to have *complete recourse* if $K_2 = \mathbb{R}^{m_0}$, that is, if the second-stage problem is feasible for any value of x. The problem has *relatively complete recourse* if $K_1 \subseteq K_2$, that is, if the second-stage problem is feasible for any value of the first-stage variables that satisfies the first-stage constraints. *Simple recourse* refers to the case when the resource matrix $W(\boldsymbol{\omega}) = I$ and the recourse constraints take the simple form $Iy_+ - Iy_- = h(\boldsymbol{\omega}) - T(\boldsymbol{\omega})x$, where I is the identity matrix, and the recourse vector y is written as $y \doteq y_+ - y_-$ with $y_+ \geqslant 0$, $y_- \geqslant 0$.

2.1.4. Deterministic equivalent formulation

We consider now the case where the random vector $\boldsymbol{\omega}$ has a discrete and finite distribution, with support $\Omega = \{\omega^1, \omega^2, \ldots, \omega^N\}$. In this case the set Ω is called a *scenario set*. Denote by p^l the probability of realization of the lth scenario ω^l. That is, for every $l = 1, 2, \ldots, N$,

$$\begin{aligned} p^l &\doteq \operatorname{Prob}\big(\boldsymbol{\omega} = \omega^l\big) \\ &= \operatorname{Prob}\big\{\big(q(y, \boldsymbol{\omega}), W(\boldsymbol{\omega}), h(\boldsymbol{\omega}), T(\boldsymbol{\omega})\big) = \big(q\big(y, \omega^l\big), W\big(\omega^l\big), h\big(\omega^l\big), T\big(\omega^l\big)\big)\big\}, \end{aligned}$$

where $p^l > 0$ for all $\omega^l \in \Omega$, and $\sum_{l=1}^N p^l = 1$.

The expected value of the second-stage optimization problem can be expressed as

$$E\big[\mathcal{Q}(x, \boldsymbol{\omega})\big] = \sum_{l=1}^{N} p^l \mathcal{Q}\big(x, \omega^l\big). \tag{8}$$

For each realization of the random vector $\omega^l \in \Omega$ a different second-stage decision is made, which is denoted by y^l. The resulting second-stage problems can then be written as:

$$\begin{aligned} &\text{Minimize} && q\big(y^l, \omega^l\big) \\ &\text{subject to} && W\big(\omega^l\big)y^l = h\big(\omega^l\big) - T\big(\omega^l\big)x, \\ & && y^l \in \mathbb{R}^{m_1}_+. \end{aligned} \tag{9}$$

Combining now (8) and (9) we reformulate the stochastic nonlinear program (7) as the following *large-scale deterministic equivalent* nonlinear program:

$$\text{Minimize} \quad f(x) + \sum_{l=1}^{N} p^l q\big(y^l, \omega^l\big) \tag{10}$$

$$\text{subject to} \quad Ax = b, \tag{11}$$

$$T\big(\omega^l\big)x + W\big(\omega^l\big)y^l = h\big(\omega^l\big) \quad \text{for all } \omega^l \in \Omega, \tag{12}$$

$$x \in \mathbb{R}^{m_0}_+, \tag{13}$$

$$y^l \in \mathbb{R}^{m_1}_+. \tag{14}$$

The constraints (11)–(14) for this deterministic equivalent program can be combined into a matrix equation with block-angular structure:

$$\begin{pmatrix} A & & & & \\ T(\omega^1) & W(\omega^1) & & & \\ T(\omega^2) & & W(\omega^2) & & \\ \vdots & & & \ddots & \\ T(\omega^N) & & & & W(\omega^N) \end{pmatrix} \begin{pmatrix} x \\ y^1 \\ y^2 \\ \vdots \\ y^N \end{pmatrix} = \begin{pmatrix} b \\ h(\omega^1) \\ h(\omega^2) \\ \vdots \\ h(\omega^N) \end{pmatrix}. \tag{15}$$

2.1.5. Multistage models

The recourse problem is not restricted to the two-stage formulation. It is possible that observations are made at T different stages and are captured in the information sets $\{\mathcal{A}_t\}_{t=1}^{T}$ with $\mathcal{A}_1 \subset \mathcal{A}_2 \subset \cdots \subset \mathcal{A}_T$. Stages correspond to time instances when some information is revealed and a decision can be made. (Note that T is a time index, while $T(\boldsymbol{\omega})$ are matrices.)

A multistage stochastic program with recourse will have a recourse problem at stage τ conditioned on the information provided by $\mathcal{A}_\tau$, which includes all information provided by the information sets $\mathcal{A}_t$, for $t = 1, 2, \ldots, \tau$. The program also anticipates the information in $\mathcal{A}_t$, for $t = \tau + 1, \ldots, T$.

Let the random vector $\boldsymbol{\omega}$ have support $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_T$, which is the product set of all individual support sets Ω_t, $t = 1, 2, \ldots, T$. $\boldsymbol{\omega}$ is written componentwise as $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_T)$. Denote the first-stage variable vector by y_0. For each stage $t =$

$1, 2, \ldots, T$, define the recourse variable vector $y_t \in \mathbb{R}^{m_t}$, the random cost function $q_t(y_t, \boldsymbol{\omega}_t)$, and the random parameters $\{T_t(\boldsymbol{\omega}_t), W_t(\boldsymbol{\omega}_t), h_t(\boldsymbol{\omega}_t) \mid \boldsymbol{\omega}_t \in \Omega_t\}$.

The multistage program, which extends the two-stage model (7), is formulated as the following nested optimization problem

$$\begin{aligned}
\text{Minimize} \quad & f(y_0) + E\Big[\operatorname*{Min}_{y_1 \in \mathbb{R}_+^{m_1}} q_1(y_1, \boldsymbol{\omega}_1) + \cdots E\Big[\operatorname*{Min}_{y_T \in \mathbb{R}_+^{m_T}} q_T(y_T, \boldsymbol{\omega}_T)\Big] \cdots\Big] \\
\text{subject to} \quad & T_1(\boldsymbol{\omega}_1) y_0 + W_1(\boldsymbol{\omega}_1) y_1 = h_1(\boldsymbol{\omega}_1), \\
& \vdots \\
& T_T(\boldsymbol{\omega}_T) y_{T-1} + W_T(\boldsymbol{\omega}_T) y_T = h_K(\boldsymbol{\omega}_T), \\
& y_0 \in \mathbb{R}_+^{m_0}.
\end{aligned} \tag{16}$$

For the case of discrete and finitely distributed probability distributions it is again possible to formulate the multistage model into a deterministic equivalent large-scale nonlinear program.

2.2. Stochastic programming model for portfolio management

Portfolio management problems can be viewed as multiperiod dynamic decision problems where transactions take place at discrete time points. At each point in time the manager has to assess the prevailing market conditions (such as prices and interest rates) and the composition of the existing portfolio. The manager also has to assess the potential future fluctuations in interest rates, prices, and cashflows. This information is incorporated into a sequence of actions of buying or selling securities, and short-term borrowing or lending. Thus, at the next point in time the portfolio manager has a seasoned portfolio and, faced with a new set of possible future movements, must incorporate the new information so that transactions can be executed.

The model specifies a sequence of investment decisions at discrete time points. Decisions are made at the beginning of each time period. The portfolio manager starts with a given portfolio and a set of scenarios about future states of the economy which she incorporates into an investment decision. The precise composition of the portfolio depends on transactions at the previous decision point and on the scenario realized in the interim. Another set of investment decisions are made that incorporate both the current status of the portfolio and new information about future scenarios.

We develop a three-stage model, with decisions made at time instances t_0, t_1, and t_2. Extension to a multistage model is straightforward. Scenarios unfold between t_0 and t_1, and then again between t_1 and t_2. A simple three-stage problem is illustrated in the *event tree* of Figure 1. An event tree shows the flow of information across time. In the example of this figure it is assumed that scenarios evolve on a binomial tree. At instance t_0 two alternative states of nature are anticipated and by instance t_1 this uncertainty is resolved. Denote these states by s_0^0 and s_0^1. At t_1 two more states are possible, s_1^0 and s_1^1. A complete path is denoted by a pair of states and such a pair is a scenario. In this example there are four scenarios from t_0 to t_2 denoted by the pairs (s_0^0, s_1^0), (s_0^0, s_1^1), (s_0^1, s_1^0),

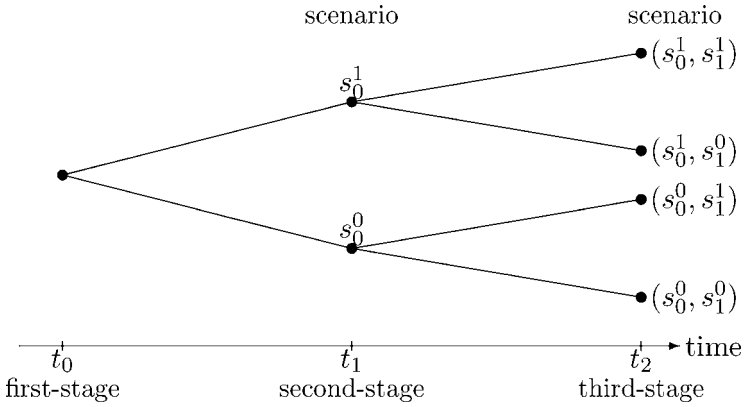


Fig. 1. Scenarios and the flow of information through time on a binomial event tree.

and (s_0^1, s_1^1). In the context of a multistage formulation introduced earlier the states s_0^0 and s_0^1 are indices of scenarios from the set Ω_1, and the states s_1^0 and s_1^1 indices of scenarios from Ω_2. The scenarios of the stochastic program are the pairs drawn from $\Omega = \Omega_1 \times \Omega_2$.

The stochastic programming model will determine an optimal decision for each state of the event tree, given the information available at that point. As there are multiple succeeding states the optimal decisions will not exploit hindsight, but they should anticipate future events.

2.2.1. Notation

The model is developed using variables to represent the buying and selling of securities, investments in the riskless asset and holdings of securities in the portfolio. Investment decisions are in dollars of face value. Some models in the literature—especially those dealing with strategic asset allocation—define decision variables in percentages of total wealth, which is usually normalized to 1 unit of the risk free asset. We define first the parameters of the model.

S_0, S_1: the index sets of states anticipated at t_0 and t_1 respectively. We use s_0 and s_1 to denote states from S_0 and S_1, respectively. Scenarios are denoted by pairs of the form (s_0, s_1), and with each scenario we associate a probability $p(s_0, s_1)$.

I: the index set of available securities or asset classes. The cardinality of I (i.e., number of available investment opportunities) is m.

c_0: the dollar amount of riskless asset available at t_0.

$b_0 \in \mathbb{R}^m$: a vector whose components denote the composition of the initial portfolio.

P_0^a, $P_0^b \in \mathbb{R}^m$: vectors of ask and bid prices respectively, at t_0. These prices are known with certainty. In order to buy an instrument the buyer has to pay the price asked by

traders, and in order to sell it the owner must be willing to accept the price at which traders are bidding.

$P_1^a(s_0)$, $P_1^b(s_0) \in \mathbb{R}^m$, for all $s_0 \in S_0$: vectors of ask and bid prices, respectively, realized at t_1. These prices depend on the state s_0.

$P_2^a(s_0, s_1)$, $P_2^b(s_0, s_1) \in \mathbb{R}^m$, for all $s_0 \in S_0$ and all $s_1 \in S_1$: vectors of ask and bid prices, respectively, realized at t_2. These prices depend on the scenario (s_0, s_1).

$\alpha_0(s_0), \alpha_1(s_0, s_1) \in \mathbb{R}^m$, for all $s_0 \in S_0$ and all $s_1 \in S_1$: vectors of *amortization factors* during the time intervals $[t_0, t_1)$ and $[t_1, t_2)$ respectively. The amortization factors indicate the fraction of outstanding face value of the securities at the end of the interval compared to the outstanding face value at the beginning of the interval. These factors capture the effects of any embedded options, such as prepayments and calls, or the effect of lapse behavior. For example, a corporate security that is called during the interval has an amortization factor of 0, and an uncalled bond has an amortization factor of 1. A mortgage security that experiences a 10 percent prepayment and that pays, through scheduled payments, an additional 5 percent of the outstanding loan has an amortization factor of 0.85. These factors depend on the scenarios.

$k_0(s_0), k_1(s_0, s_1) \in \mathbb{R}^m$, for all $s_0 \in S_0$, and all $s_1 \in S_1$: vectors of *cash accrual factors* during the intervals $[t_0, t_1)$ and $[t_1, t_2)$ respectively. These factors indicate cash generated during the interval, per unit face value of the security, due to scheduled payments and exercise of the embedded options, accounting for accrued interest. For example, a corporate security that is called at the beginning of a one-year interval, in a 10 percent interest rate environment, will have a cash accrual factor of 1.10. These factors depend on the scenarios.

$\rho_0(s_0)$, $\rho_1(s_0, s_1)$: short-term riskless reinvestment rates during the intervals $[t_0, t_1)$ and $[t_1, t_2)$ respectively. These rates depend on the scenarios.

$L_1(s_0)$, $L_2(s_0, s_1)$: liability payments at t_1 and t_2 respectively. Liabilities may depend on the scenarios as discussed in Section 3.1.

Now let us define decision variables. We have four distinct decisions at each point in time: how much of each security to buy, sell, or hold in the portfolio, and how much to invest in the riskless asset. All variables are constrained to be nonnegative.

First-stage variables at t_0:

$x_0 \in \mathbb{R}^m$: the components of the vector denote the face value of each security bought.

$y_0 \in \mathbb{R}^m$: denotes, componentwise, the face value of each security sold.

$z_0 \in \mathbb{R}^m$: denotes, componentwise, the face value of each security held in the portfolio.

v_0^+: the dollar amount invested in the riskless asset.

Second-stage variables at t_1 for each state $s_0 \in S_0$:

$x_1(s_0) \in \mathbb{R}^m$: denotes the vector of the face values of each security bought.

$y_1(s_0) \in \mathbb{R}^m$: denotes the vector of the face values of each security sold.

$z_1(s_0) \in \mathbb{R}^m$: denotes the vector of the face values of each security held in the portfolio.

$v_1^+(s_0)$: the dollar amount invested in the riskless asset.

Third-stage variables at t_2 for each scenario (s_0, s_1) such that $s_0 \in S_0$ and $s_1 \in S_1$:

$x_2(s_0, s_1) \in \mathbb{R}^m$: denotes the vector of the face values of each security bought.

$y_2(s_0, s_1) \in \mathbb{R}^m$: denotes the vector of the face values of each security sold.
$z_2(s_0, s_1) \in \mathbb{R}^m$: denotes the vector of the face values of each security held in the portfolio.
$v_2^+(s_0, s_1)$: the dollar amount invested in the riskless asset.

2.2.2. *Model formulation*

There are two basic constraints in stochastic programming models for portfolio optimization. One expresses cashflow accounting for the riskless asset, and the other is an inventory balance equation for each asset class or each security at all time periods.

2.2.2.1. First-stage constraints: At the first stage (i.e., at time t_0) all prices are known with certainty. The *cashflow accounting equation* specifies that the original endowment in the riskless asset, plus any proceeds from liquidating part of the existing portfolio, equal the amount invested in the purchase of new securities plus the amount invested in the riskless asset, i.e.,

$$c_0 + \sum_{i=1}^{m} P_{0i}^b y_{0i} = \sum_{i=1}^{m} P_{0i}^a x_{0i} + v_0^+. \tag{17}$$

For each asset class in the portfolio we have an *inventory balance constraint*:

$$b_{0i} + x_{0i} = y_{0i} + z_{0i} \quad \text{for all } i \in I. \tag{18}$$

2.2.2.2. Second-stage constraints: Decisions made at the second stage (i.e., at time t_1) depend on the state s_0 realized during the interval $[t_0, t_1)$. Hence, we have one constraint for each state. These decisions also depend on the investment decisions made at the first stage, i.e., at t_0.

Cashflow accounting ensures that the amount invested in the purchase of new securities and the riskless asset is equal to the income generated by the existing portfolio during the holding period, plus any cash generated from sales, less the liability payments. There is one constraint for each state:

$$\begin{aligned}\big(1 + \rho_0(s_0)\big)v_0^+ + \sum_{i=1}^{m} k_{0i}(s_0) z_{0i} + \sum_{i=1}^{m} P_{1i}^b(s_0) y_{1i}(s_0) \\ = v_1^+(s_0) + \sum_{i=1}^{m} P_{1i}^a(s_0) x_{1i}(s_0) + L_1(s_0), \quad \text{for all } s_0 \in S_0.\end{aligned} \tag{19}$$

This constraint allows investment in the riskless asset (variable v_1^+) but not borrowing. Borrowing can be incorporated in this equation by introducing a new variable v^-. Borrowing will contribute to the cash inflow (left-hand side of the equation above) but borrowing from previous time periods must be paid back, with proper interest, at subsequent periods. This will increase the cash outflows (right-hand side of the equation

above). The cashflow accounting equation with borrowing and reinvestment at each state $s_0 \in S_0$ is written as follows:

$$\begin{aligned}&\big(1+\rho_0(s_0)\big)v_0^+ + \sum_{i=1}^{m} k_{0i}(s_0)z_{0i} + \sum_{i=1}^{m} P_{1i}^b(s_0)y_{1i}(s_0) + v_1^-(s_0)\\ &\quad = v_1^+(s_0) + \sum_{i=1}^{m} P_{1i}^a(s_0)x_{1i}(s_0) + L_1(s_0) + v_0^-\big(1+\rho_0(s_0)+\delta\big)\end{aligned} \tag{20}$$

where δ is the spread between borrowing and lending rates.

Inventory balance equations constrain the amount of each security sold or remaining in the portfolio to be equal to the outstanding amount of face value at the end of the first period, plus any amount purchased at the beginning of the second stage. There is one constraint for each security and for each state:

$$\alpha_{0i}(s_0)z_{0i} + x_{1i}(s_0) = y_{1i}(s_0) + z_{1i}(s_0), \quad \text{for all } i \in I,\ s_0 \in S_0. \tag{21}$$

2.2.2.3. Third-stage constraints: Decisions made at the third stage (i.e., at time t_2) depend on the scenario (s_0, s_1) realized during the period $[t_1, t_2)$ and on the decisions made at t_1. The constraints are similar to those of the second stage. The cashflow accounting equation, without borrowing, is

$$\begin{aligned}&\big(1+\rho_1(s_0,s_1)\big)v_1^+(s_0) + \sum_{i=1}^{m} k_{1i}(s_0,s_1)z_{1i}(s_0) + \sum_{i=1}^{m} P_{2i}^b(s_0,s_1)y_{2i}(s_0,s_1)\\ &\quad = v_2^+(s_0,s_1) + \sum_{i=1}^{m} P_{2i}^a(s_0,s_1)x_{2i}(s_0,s_1) + L_2(s_0,s_1),\end{aligned} \tag{22}$$

for all scenarios (s_0, s_1) such that $s_0 \in S_0$ and $s_1 \in S_1$.

The inventory balance equation is:

$$\alpha_{1i}(s_0,s_1)z_{1i}(s_0) + x_{2i}(s_0,s_1) = y_{2i}(s_0,s_1) + z_{2i}(s_0,s_1), \tag{23}$$

for all $i \in I$, and all scenarios (s_0, s_1) such that $s_0 \in S_0$ and $s_1 \in S_1$.

2.2.2.4. Other conditions: At each stage of the stochastic program we have formulated two sets of constraints: cashflow accounting and the inventory balance. Depending on the application at hand other conditions may need to be modelled as constraints. The general setup with the variables as defined here is usually adequate for formulating additional constraints. We discuss several examples of conditions that appear in practice. Some applications require multiple cash accounts. For instance, international portfolio management requires different handling of cash in different currencies when exchange rates are hedged (Consiglio and Zenios, 2001; Topaloglou, Vladimirou and Zenios, 2002). Deposits from different product lines may be held in separate accounts when regulators apply different rules for different sources. This is the case for Japanese saving type insurance policies that are treated differently than conventional policies (Carino

et al., 1994). Other conditions may include limits on the position in a given asset class. For instance, the allowable exposure of Italian insurers to corporate bonds or international Government bonds limited by regulators, see the chapter by Consiglio, Cocco and Zenios in Volume 2 of this Handbook. Investments in "tokkin" funds by Japanese insurers may not exceed seven percent of the total assets (Carino et al., 1994). Taxes must be computed distinguishing income return from price return, and this requirement can be formulated using the sales variables (y) to model income return, and the inventory variables (z) to model price return. Leverage restrictions may be imposed by regulators requiring the calculation of the ratio of debt to equity in funding liabilities. These, and several other conditions, may be imposed to the basic constraints formulated above.

2.2.2.5. Objective function: The objective function maximizes the expected utility of terminal wealth. In order to measure terminal wealth all securities in the portfolio are marked-to-market. This approach is in agreement with U.S. Federal Accounting Standards Board (FASB) regulations that require reporting portfolio market and book values. The composition of the portfolio and its market value depend on the scenarios (s_0, s_1). The objective of the portfolio optimization model is

$$\text{Maximize} \quad \sum_{(s_0,s_1)\in S_0\times S_1} p(s_0, s_1)\mathcal{U}\big(W(s_0, s_1)\big),$$

where $p(s_0, s_1)$ is the probability associated with scenario (s_0, s_1); $W(s_0, s_1)$ denotes terminal wealth; and $\mathcal{U}$ denotes the utility function. Terminal wealth is given by

$$W(s_0, s_1) \doteq v_2^+(s_0, s_1) + \sum_{i=1}^{m} P_{2i}^b(s_0, s_1)z_{2i}(s_0, s_1). \tag{24}$$

This is not by any means a standard objective function, although it is the one in agreement with the literature on discrete multi-period models (Mossin, 1968; Samuelson, 1969; Hakansson, 1970). Other choices may be more appropriate for some applications. For instance, Dert (1995) minimizes the expected cost of funding a defined benefits pension fund. Carino et al. (1994) consider a multicriteria objective function that maximizes terminal wealth net of expected shortfalls. Consiglio, Cocco and Zenios (2001) consider the maximization of return-on-equity to shareholders as a proxy for shareholder value. For indexed funds the objective function is a measure of deviation of portfolio returns from the target index. Quite often only downside deviations are minimized. The case studies in this handbook by Ziemba, Mulvey and Thomas, and Høyland and Wallace discuss different objective functions as well.

In general creating an objective function for investors over long time horizons is a poorly understood task. First, temporal considerations trading short return versus long-term goals must be estimated. Second, uncertainty over extended time periods complicates the decision making process by creating potential regret. Reconciling the choice of an objective function with accepted theories on investor preferences and utility functions is an important step of the modelling process.

3. Scenario generation and tree construction

An important issue for successful applications of stochastic programming models is the construction of *event trees* with asset and liability returns. As a first step, a return generating process for the assets and relevant economic factors has to specified. This task can be quite complicated, as many economic factors affect the assets and liabilities of a large firm, pension fund or financial institution. Second, the liability values have to be estimated with appropriate rules taking into account actuarial risks, pension or social security fund provisions, and other relevant factors for the institution's line of business. We describe some of the simulation systems that have been proposed in the literature to handle the complicated task of scenario generation for ALM (Zenios, 1991, 1995; Mulvey and Zenios, 1994; Mulvey, 1996; Boender, 1997; Carino et al., 1994; Carino, Myers and Ziemba, 1998, amongst others).

A third important step is the representation of the underlying return generating process in the event tree that is used as input in a multi-stage stochastic programming model for ALM. The return distributions often have to be represented with a small number of nodes in the event tree, as otherwise the computational effort for solving the stochastic program would explode. Clearly, a small number of nodes in the event tree for describing the return distribution might lead to approximation error. An important question is the extent to which the approximation error in the event tree will bias the optimal solutions of the model. Moreover, event trees for ALM models with options and interest rate dependent securities require special attention to preclude arbitrage opportunities. We discuss research on these important issues, including Carino et al. (1994), Shtilman and Zenios (1993), Klaassen (1997, 1998), Pflug (2001), Kouwenberg (2001), Høyland and Wallace (2001) and Gondzio, Kouwenberg and Vorst (2003).

Throughout this section we will use the word "scenario" for one possible series of future asset returns and economic variables, which is typically generated by a simulation system. We will use "event tree" to denote a set of scenarios with a tree structure as in Figure 1, which is used as input for a multi-stage stochastic programming model. We make an important distinction between the processes of "scenario generation" and "tree construction". According to our definition the aim of scenario generation is to create a large set of scenarios as a forecast of the many possible future states of the world that might occur. The aim of tree construction is to build an event tree for a multi-stage stochastic programming model, such that the size of the stochastic program remains tractable, while still representing the underlying distribution without large approximation errors.

3.1. Scenarios for the liabilities

Any asset liability management model requires a projection of the future value of the liabilities. The liabilities typically represent the discounted expected value of the future obligatory payments by the financial institution or firm. Examples include liabilities

resulting from bank deposits, pension fund or social security liabilities due to future benefit payments, and liabilities resulting from the sale of insurance contracts. Each firm and financial institution typically has its own unique set of liabilities. Hence we can not provide a general recipe for calculating the value of the liabilities. For pension funds and insurance companies actuarial methods can be very important, while other financial institutions might require financial economic valuation models (Embrechts, 2000).

The liabilities of pension funds and insurance companies usually consist of a large number of individual contracts, and the development of the total liability value is influenced by multiple sources of uncertainty. As this setting frustrates mathematical analysis, simulation is an important approach for ALM applications with a complex liability structure. A simulation model must be able to capture the complex interactions between the state of the economy, the financial markets, security prices and the value of the liabilities. Figure 2 illustrates a hierarchy of simulation models that capture these interactions. User intervention is an important part of the process as some effects can not be captured by simulation models. These are sometimes called "Gorbachev effects" in reference to changes brought about by events that could not be anticipated in any simulation model.

The development of sophisticated multi-period simulation models for asset liability management is already reported in the early eighties by Goldstein and Markowitz (1982), Kingsland (1982) and Winklevoss (1982). Simulation models try to replicate the composition and development of the liability structure as closely as possible in order to increase the accuracy of liability estimates. Macroeconomic variables and actuarial predictions drive the liability side, whereas the economic variables drive the financial markets and determine security prices on the asset side.

We discuss briefly two examples to illustrate the issues. A simulation system for Dutch pension funds developed by Boender (1997) first simulates the future status of a large group of fund participants, according to assumed mortality rates, retirement rates, job termination rates and career promotion probabilities. As a second step, values of the future wage growth are simulated, as this is an important factor for determining the pension payments in the long run. Combining the simulated status of the individual participants with the simulated wage growth, the future value of the liabilities in the scenarios is calculated as a discounted expected value of the pension payments. Established actuarial rules apply for the valuation once the economic and financial scenarios have been generated. Regulatory authorities check the solvency of pension funds by comparing the liabilities to the asset value of the fund, and hence accurate projections of the future liabilities are important for financial planning.

A different approach is taken by financial institutions, such as banks and money management firms, dealing primarily with assets and liabilities that are mostly influenced by changes in interest rates. The same is true with some products offered by insurers. The nature of their business gives rise to another important class of asset liability management problems. For instance, the bank ALM problem described by Kusy and Ziemba (1986) has fixed term deposits as the main liability. Other examples include government

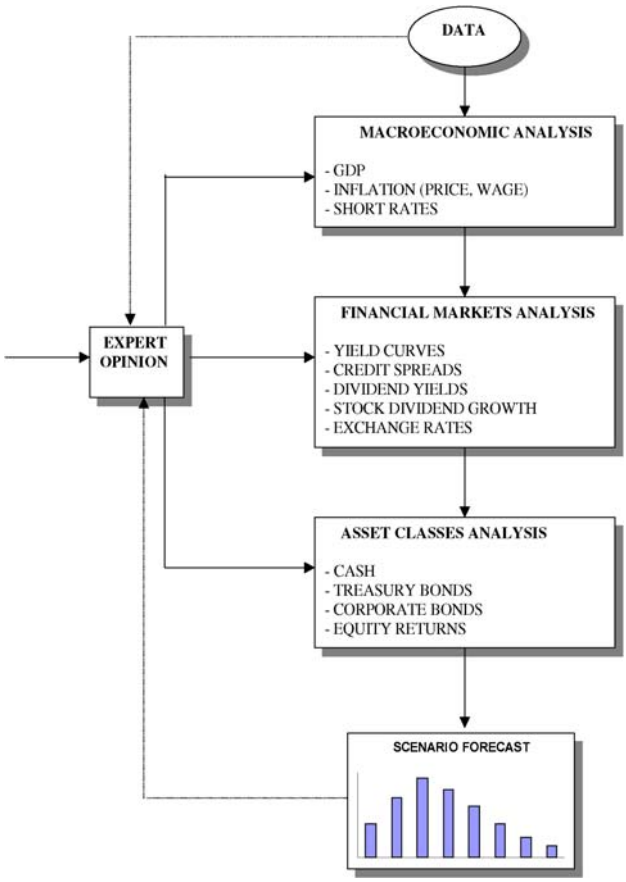


Fig. 2. Hierarchy of simulation models for the generation of scenarios for asset and liability management.

agencies such as Fannie Mae and Freddie Mac, that fund the purchase of mortgages by issuing debt (Holmer, 1994; Zenios, 1995), fixed-income money managers (Golub et al., 1995), insurance companies that offer combined insurance and investment products (Asay, Bouyoucos and Marciano, 1993 and the chapter by Consiglio, Cocco and Zenios in Volume 2 of this Handbook). The liabilities of defined contribution pension plans in the US are also mainly affected by interest rate changes, as fixed payments in the future are discounted with the current market interest rate. For this class of ALM problems scenario generation methods focus on the simulation of risk free rates and other key financial primitives, such as credit spreads, liquidity premia, prepayment or lapse. Defined benefits funds and social security, on the other hand, fall under the class of models of the previous paragraph as the benefits depend, through some regulatory formula, on economic indicators such as inflation or wage growth.

Methods for generating scenarios for asset liability management problems that mainly involve interest rate dependent securities are described in the above references. First, a lattice for the short term interest rate is constructed using a model such as the one suggested by Black, Derman and Toy (1990) or Hull and White (1990). An important property is that the prices of treasury bonds computed with this lattice are consistent with the initial yield curve. Second, the prices of other relevant interest rate dependent securities (such as mortgage backed securities, single-premium deferred annuities, callable bonds) and the value of the liabilities are added to the lattice by applying financial economic valuation rules and simulations for other factors such as prepayments or lapse. Finally, consistent scenarios of interest rate movements, fixed income prices and liability values can be constructed by sampling paths from the lattice. This methodology generates price scenarios under the risk-neutral probability measure. For short horizons the risk neutral and the objective measure are indistinguishable and lattice-based scenario generation is valid. For long horizons the methodology breaks down, except for some problems involving index replication, and a risk premium must be properly estimated and incorporated in the valuation stage.

An important distinction must be made in this class of models between state-dependent and path-dependent instruments. In the former case the price of an instrument is uniquely determined at each state of the lattice. In the later case the prices depend on the path that leads to a given state. While the number of states is a polynomial function of the number of steps, the number of paths grows exponentially. A 360-step lattice of monthly steps over 30-years has $360^2/2$ states but 2^{360} paths. High-performance computations may be needed for the simulation of path-dependent securities (Cagan, Carriero and Zenios, 1993).

3.2. Scenarios for economic factors and asset returns

Asset liability management applications typically require simulation systems that integrate the asset prices with the liability values. This integration is crucial as the assets and liabilities are often affected by the same underlying economic factors. For example, in pension fund simulations wage growth and inflation are crucial factors for the value of the liabilities, and these factors are also associated with the long run returns on stocks and bonds (Boender, 1997). In fixed income ALM applications for money management the short term interest rate is driving the returns on both assets and liabilities (Zenios, 1995). We echo the first chapter of this handbook that the integration of assets and liabilities is crucial for successful ALM applications at an enterprise-wide level. The integration starts with the consistent simulation of future scenarios for both sides of the balance sheet. As the liabilities are often unique and different in each ALM application, we will from now on concentrate on generating scenarios for economic factors and asset returns. Values of the liabilities may be added to the economic scenarios with a consistent method following actuarial practices or standard financial valuation tools (Embrechts, 2000).

Perhaps the most complete instantiation of the framework illustrated in Figure 2 is the scenario generation system developed by the company Towers Perrin for pension management problems (Mulvey, 1996). The economic forecasting system consists of a linked set of modules that generate scenarios for different economic factors and asset returns. At the highest level of the system, the Treasury yield curve is modelled by a two-factor model based on Brennan and Schwartz (1982). Other models could have been used here as well, perhaps accounting for market shocks. Based on the scenarios for the short and consol rates, other modules generate forecasts of the price inflation, bond returns and the dividend yield on stocks. After the return on a major stock index (e.g., the S&P 500) has been generated conditional on the dividend yield, the return on corporate bonds and small cap stocks are derived at the lowest level of the system. The cascade design of the Towers Perrin scenario generation system limits the number of coefficients that have to be estimated with the available data and leads to consistent forecasts for the returns on a large number of assets.

Other models generating asset returns are described in Brennan, Schwartz and Lagnado (1997) for strategic asset allocation, Carino et al. (1994), Carino, Myers and Ziemba (1998) for an insurance company, Consiglio, Cocco and Zenios (this volume) for minimum guarantee products, amongst others. These approaches do not model in detail the economic conditions. Given the interest-rate dependence of the liabilities in these studies, and the strategic decisions they address, this omission was not significant. However, a more general model is needed for defined benefits pension funds, social security funds, long term insurance products and so on.

Once the model for generating scenarios has been specified, the coefficients have to be calibrated in order to produce plausible values for the returns. For example, the Towers Perrin system consists of a number of diffusions for the key economic factors such as the interest rate and the dividend yield. The coefficients of these diffusions have to be estimated: one can apply a pragmatic approach that matches historical summary statistics and expert opinions (Høyland and Wallace, 2001; Mulvey, Gould and Morgan, 2000) or traditional econometric methods for discrete-time models (Green, 1990; Hamilton, 1994) and for continuous time-models (Duffie and Singleton, 1993; Hansen and Scheinkman, 1995). ALM applications with fixed income securities, such as mortgage assets, are often based on interest rate lattice models. These require the calibration of a lattice that perfectly matches the current yield curve of treasury bills and bonds. Proper calibration ensures that the coefficients of a model are consistent with historical data or current prices (Black, Derman and Toy, 1990; Hull and White, 1990). As scenarios are projections of the future, the users of ALM models can of course adjust the estimated coefficients in order to incorporate their own views about the economy and the asset markets (Koskosides and Duarte, 1997). Sometimes stress scenarios are incorporated in response to requirements by the supervisory authorities or to satisfy corporate safeguards.

3.3. Methods for generating scenarios

In this subsection we describe three specific methods for generating asset return scenarios with more detail: (i) bootstrapping historical data, (ii) statistical modelling with the Value-at-Risk approach, and (iii) modelling economic factors and asset returns with vector autoregressive models.

3.3.1. Bootstrapping historical data

The simplest approach for generating scenarios using only the available data without any mathematical modelling is to bootstrap a set of historical records. Each scenario is a sample of returns of the assets obtained by sampling returns that were observed in the past. Dates from the available historical records are selected randomly and for each date in the sample we read the returns of all asset classes or risk factors during the month prior to that date. This are scenarios of *monthly* returns. If we want to generate scenarios of returns for a long horizon—say 1 year—we sample 12 monthly returns from different points in time. The compounded return of the sampled series is the 1-year return. Note that with this approach the correlations among asset classes are preserved.

3.3.2. Statistical models from the Value-at-Risk literature

Time series analysis of historical data can be used to estimate volatilities and correlation matrices among asset classes of interest. Riskmetrics (1996) has become an industry standard in this respect. These correlation matrices are used to measure risk exposure of a position through the Value-at-Risk (VaR) methodology.

Denote the random variables by the K-dimensional random vector $\boldsymbol{\omega}$. The dimension of $\boldsymbol{\omega}$ is equal to the number of risk factors we want to model. Assuming that the random variables are jointly normally distributed we can define their probability density function of $\boldsymbol{\omega}$ by

$$f(\boldsymbol{\omega}) = (2\pi)^{-p/2}|Q|^{-1/2}\exp\left[-\frac{1}{2}(\boldsymbol{\omega}-\bar{\boldsymbol{\omega}})'Q^{-1}(\boldsymbol{\omega}-\bar{\boldsymbol{\omega}})\right], \tag{25}$$

where $\bar{\boldsymbol{\omega}}$ is the expected value of $\boldsymbol{\omega}$ and Q is the covariance matrix and they can be calculated from historical data. (It is typically the case in financial time series to assume that the logarithms of the changes of the random variables have the above probability density function, so that the variables themselves follow a lognormal distribution.)

Once the parameters of the multivariate normal distribution are estimated we can use it in Monte Carlo simulations, using either the standard Cholesky factorization approach (see, e.g., RiskMetrics, 1996, Chapter 7) or scenario generation procedures based on principal component analysis discussed in Jamshidian and Zhu (1997).

The simulation can be applied repeatedly at different states of an event tree. However, we may want to condition the generated random values on the values obtained by some of the random variables. For instance, users may have views on some of the variables,

or a more detailed model may be used in the simulation hierarchy to estimate some of the variables. This information can be incorporated when sampling the multivariate distribution.

The conditional sampling of multivariate normal variables proceeds as follows. Variable $\boldsymbol{\omega}$ is partitioned into two subvectors $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$, where ω_1 is the vector of dimension K_1 of random variables for which some additional information is available and $\boldsymbol{\omega}_2$ is the vector of dimension $K_2 = K - K_1$ of the remaining variables. The expected value vector and covariance matrix are partitioned similarly as

$$\bar{\boldsymbol{\omega}} = \begin{bmatrix} \bar{\boldsymbol{\omega}}_1 \\ \bar{\boldsymbol{\omega}}_2 \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}. \tag{26}$$

The marginal probability density function of $\boldsymbol{\omega}_2$ given $\boldsymbol{\omega}_1 = \boldsymbol{\omega}_1^*$ is given by

$$f\left(\boldsymbol{\omega}_2 \mid \boldsymbol{\omega}_1 = \boldsymbol{\omega}_1^*\right) = (2\pi)^{-p_2/2} |Q_{22.1}|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{\omega}_2 - \bar{\boldsymbol{\omega}}_{2.1})' Q_{22.1}^{-1} (\boldsymbol{\omega}_2 - \bar{\boldsymbol{\omega}}_{2.1})\right], \tag{27}$$

where the conditional expected value and covariance matrix are given by

$$\bar{\boldsymbol{\omega}}_{2.1}\left(\boldsymbol{\omega}_1^*\right) = \left(\bar{\boldsymbol{\omega}}_2 - Q_{21} Q_{11}^{-1} \mu_1\right) + Q_{21} Q_{11}^{-1} \boldsymbol{\omega}_1^*, \tag{28}$$

and

$$Q_{22.1} = Q_{22} - Q_{21} Q_{11}^{-1} Q_{12}, \tag{29}$$

respectively. Scenarios of $\boldsymbol{\omega}_2$ for period t conditioned on values of $\boldsymbol{\omega}_1$ given by ω_1^* can be generated from the multivariate normal variables from (27) through the expression

$$\boldsymbol{\omega}_{2i}^t = \boldsymbol{\omega}_{2i}^0 \exp\left[\sigma_i \sqrt{t} \boldsymbol{\omega}_{2i}\right],$$

where $\boldsymbol{\omega}_{2i}^0$ is today's value and σ_i is the single-period volatility of the ith component of the random variable ω_2.

Consiglio and Zenios (2001) use the Riskmetrics methodology *in conjunction* with discrete lattice models to generate joint scenarios of term-structure and exchange rates. Interest rate differentials among two countries are key determinants of the exchange rate between the currencies. Hence, exchange rate scenarios are conditioned on the interest rates of the two currencies, the base currency and the foreign currency. The standard assumption applies that the logarithms of the ratios of exchange rates at period t to period $t-1$, and the logarithms of the ratios of spot interest rates at period t to period $t-1$ follow a multivariate normal distribution. Daily and weekly rates do not follow normal distributions but there is lack of empirical evidence against normality for monthly data such as those used by Consiglio and Zenios.

Figure 3 illustrates the conditional probabilities for exchange rate scenarios. On the same figure we plot the exchange rate that was realized *ex post* on the date for which the scenarios were estimated. Note that the same exchange rate value may be obtained for various different scenarios of interest rates and samples drawn from (27); the figure plots several points with the same exchange rate value but different conditional probabilities.

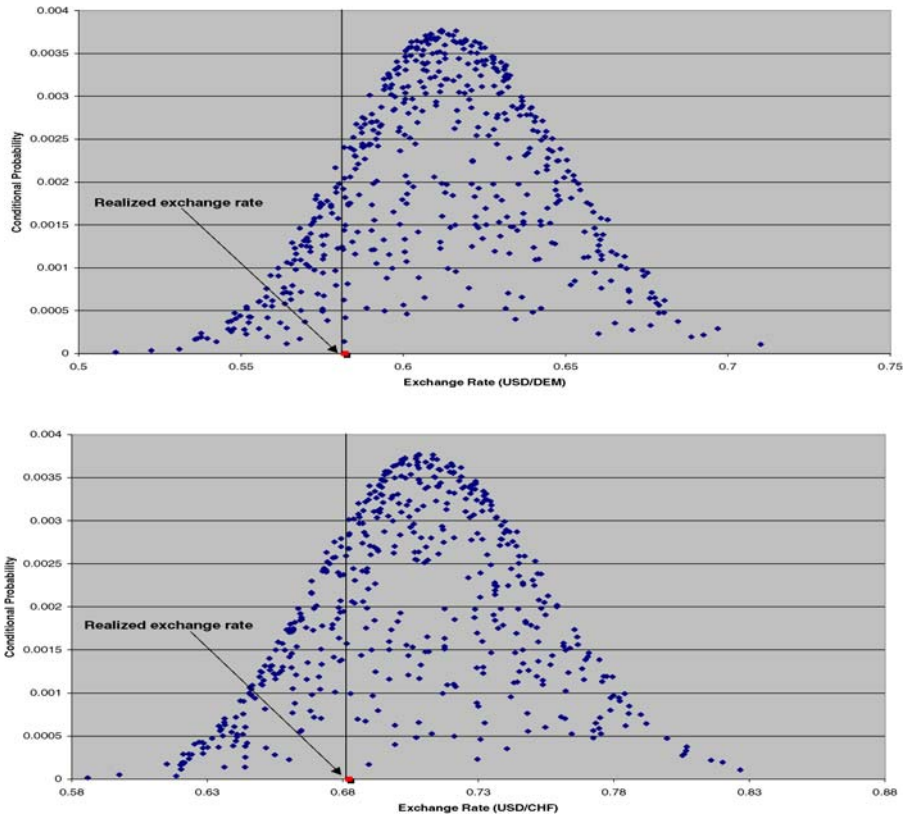


Fig. 3. Exchange rate scenarios and their conditional probabilities for the DEM and CHF against the USD.

3.3.3. Modelling economic factors and asset returns for a pension fund

To illustrate the calibration of a system of equations for economic factor values and asset returns, we consider an ALM simulation system for Dutch pension funds as an example (see Boender (1997)). As the scope of ALM systems for Dutch pension funds is often limited to long term strategic decisions, the investment model only considers a small set of broad asset classes: deposits, bonds, real estate and stocks. Apart from the returns on these assets, each scenario should contain information about future wage growth in order to calculate the future values of the pension liabilities.

To generate asset returns and the wage growth rate a vector autoregressive model is applied by Boender (1997):

$$r_t = c + V r_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, Q), \ t = 1, 2, \ldots, T, \tag{30}$$

$$r_{it} = \ln(1 + R_{it}), \quad i = 1, 2, \ldots, m, \ t = 1, 2, \ldots, T, \tag{31}$$

where m is the number of time series, R_{it} is the discrete rate of change of variable i in year t, r_t is an m-dimensional vector of continuously compounded rates, c is a m-dimensional vector of coefficients, V is an $m \times m$ matrix of coefficients, ε_t is the m-dimensional vector of error terms and Q is the $m \times m$ covariance matrix. To estimate the coefficients of the model we use yearly data on the asset returns and the general wage increase in the Netherlands from 1956 to 1994. Table 1 displays the descriptive statistics of the data and Table 2 displays the correlation matrix.

The specification of the vector autoregressive model should be chosen carefully. Although some inter-temporal relationships between the returns might be weakly significant based on historical data, that does not imply that these relationships are also useful for generating scenarios for a financial optimization model with a long time horizon. To avoid any problems with unstable and spurious predictability of returns, we do not use lagged variables for explaining the returns of bonds, real estate, and stocks in the vector autoregressive model (Table 3). The timeseries of the return on deposits and

Table 1
Statistics of time series for asset returns, 1956–1994

	Mean	St. dev.	Skewness	Kurtosis
Wages	0.061	0.044	0.434	2.169
Deposits	0.055	0.025	0.286	2.430
Bonds	0.061	0.063	0.247	3.131
Real estate	0.081	0.112	−0.492	7.027
Stocks	0.102	0.170	0.096	2.492

Wages is the rate of change of the Dutch general wage level. The time series for deposits is based on the average of the 3-month borrowing rate for government agencies. In each year a premium of 0.5% has been subtracted, because the pension fund will have to lend cash to commercial banks. The asset class of bonds represents the total return of a roll-over investment in long term Dutch government bonds. Real estate consists of total returns of the property fund Rodamco. Stocks is the total return of the internationally diversified mutual fund Robeco. All time series were provided by Ortec Consultants.

Table 2
Correlations of asset classes, annually 1956–1994

	Wages	Deposits	Bonds	Real estate	Stocks
Wages	1				
Deposits	−0.059	1			
Bonds	−0.127	0.259	1		
Real estate	0.162	−0.053	0.360	1	
Stocks	−0.296	−0.157	0.379	0.326	1

the increase of the wage level on the other hand are known to have some memory, so we model them by a first order autoregressive process.

There are many ways to estimate vector autoregressive models, see, e.g., Judge et al. (1988). Table 3 shows the coefficients estimated by the method of iterative weighted least squares, using the econometric software Eviews; Table 4 displays the estimated correlation matrix of the residuals. We would like to point out that the average return on bonds is rather low in our sample of data, due to two outliers in 1956 and 1957. In order to generate plausible future bond returns, we choose to increase the coefficient of the bond returns in the vector autoregressive model by 1%. For the purpose of generating scenarios we can adjust the coefficients based on subjective expectations, as historical data is not necessarily representative for the future. User intervention may be required as illustrated in Figure 2 and discussed in Koskosides and Duarte (1997).

Finally, scenarios of asset returns for a financial planning model can be constructed by sampling from the error distribution of the vector autoregressive model, given the estimated (and possibly user-adjusted) coefficients of Table 3. After the vector autore-

Table 3
Estimated coefficients of the vector autoregressive model

$r_{\text{wages},t}$	$= 0.018$ (2.058)	$+ 0.693\, r_{\text{wages},t-1} + \varepsilon_{1t}$ (5.789)	$\sigma_1 = 0.030$
$r_{\text{deposits},t}$	$= 0.020$ (2.865)	$+ 0.644\, r_{\text{deposits},t-1} + \varepsilon_{2t}$ (5.448)	$\sigma_2 = 0.017$
$r_{\text{bonds},t}$	$= 0.058$ (6.241)	$+ \varepsilon_{3t}$	$\sigma_3 = 0.060$
$r_{\text{realest},t}$	$= 0.072$ (4.146)	$+ \varepsilon_{4t}$	$\sigma_4 = 0.112$
$r_{\text{stocks},t}$	$= 0.086$ (3.454)	$+ \varepsilon_{5t}$	$\sigma_5 = 0.159$

Estimated with iterative weighted least squares using annual data for the period 1956–1994, t-statistics in parenthesis. σ_i denotes the mean square error (standard deviation of the residuals) for each asset return equation $i = 1, 2, \ldots, 5$.

Table 4
Residual correlations of vector autoregressive model

	Wages	Deposits	Bonds	Real estate	Stocks
Wages	1				
Deposits	0.227	1			
Bonds	−0.152	−0.268	1		
Real estate	−0.008	−0.179	0.343	1	
Stocks	−0.389	−0.516	0.383	0.331	1

gressive model has been used to generate scenarios of asset returns and wage growth, the liability values can be added to each scenario in a consistent manner by applying appropriate actuarial rules or financial valuation principles (Embrechts, 2000).

3.4. Constructing event trees

A stochastic programming model is based upon an event tree for the key random variables (Figure 1). Each node of the event tree has multiple successors, in order to model the process of information being revealed progressively through time. The stochastic programming approach will determine an optimal decision for each node of the event tree, given the information available at that point. As there are multiple succeeding nodes the optimal decisions will be determined without exploiting hindsight. If a stochastic programming model is formulated then the optimal policy will be tailor-made to fit the condition of the financial institution and the economy in each node, while anticipating the optimal adjustment of the policy later on as the tree evolves and more information is revealed.

A key issue for the successful application of stochastic programming in ALM is the construction of event trees with asset and liability returns from the scenarios generated by the processes discussed in Sections 3.1–3.3. The underlying return distributions have to be discretized with a small number of nodes in the event tree, otherwise the computational effort for solving a multi-stage stochastic programming model can easily explode. Clearly, a small number of nodes describing the return distribution at every stage of the event tree might lead to some approximation error. An important question is to which extent the approximation error in the event tree will bias the optimal solutions of the model.

In this section we will consider three different methods to construct event trees for stochastic programming models: (i) random sampling, (ii) adjusted random sampling, and (iii) tree fitting. In order to compare these methods we will apply them to construct trees with asset and liability returns for the estimated vector autoregressive model of Table 3. However these methods are applicable to any scenario generation model such as the Towers Perrin model, the lattice-based models, or the other models discussed in Section 3.3. We will also demonstrate the importance of constructing proper event trees by solving a multi-stage stochastic programming model for a Dutch pension fund using these event trees as input.

3.4.1. Random sampling

As a first method for constructing event trees we consider random sampling from the error distribution of the vector autoregressive model. Given the estimated coefficients and the estimated covariance matrix of the vector autoregressive model, we can draw one random vector of yearly returns for bonds, real estate, stocks, deposits and wage growth. If we would like to construct an event tree with ten nodes after one year (we assume that the duration of each stage is one year), we can simply repeat this procedure

ten times, sampling independent vectors of returns for each node. The nodes at stage two in the event tree can also be sampled randomly, however the conditional distribution from stage one to stage two depends on the outcomes at the first stage. For example, wage growth follows an autoregressive process, so the expected wage growth from year one to year two depends on the realized wage growth rate in the previous period. The sampled wage growth data at stage one will be input to the right-hand side of the wage equation (Table 3) when sampling error terms at stage two.

We create an entire event tree for the stochastic program by applying random sampling recursively, from stage to stage, while adjusting the conditional expectations of wage growth and deposits in each node based on previous outcomes. Applying random sampling to construct an event tree consisting of five (yearly) stages and a branching structure of 1-10-6-6-4-4 we obtain a tree with $10 \times 6 \times 6 \times 4 \times 4 = 5760$ nodes at time period 5.

As an example, we solve the stochastic programming model for Dutch pension funds of Kouwenberg (2001) with this randomly sampled event tree as input. The model minimizes the average contribution rate paid by the sponsor of the pension plan, while taking into account the downside risk of deficits and the state of the fund at the planning horizon, based on the pension fund model of Dert (1995). The decision variables of the model are the contribution rate and the strategic asset mix, consisting of bonds, stocks and real estate (see Table 3). Borrowing and lending are not allowed in the model, in order to limit the number of decision variables. The asset weights are restricted to be nonnegative in order to prevent short selling. Transaction costs of 1% are imposed both on buying and on selling of assets.

Table 5 contains information about the optimal solution of the stochastic programming model. The initial asset mix—first row of the table—is the part of the solution that could be implemented. In this case the model recommends that the pension fund invests 100% of its endowment in stocks. The mean absolute change of the asset weights

Table 5
Solution of stochastic programming model: random sampling

Initial asset mix	Bonds	Real	Stocks
	0.00	0.00	1.00
Portfolio turnover	Bonds	Real	Stocks
	0.31	0.37	0.42
Objective value	0.2294		
Mean contribution rate	0.0126		
Downside risk	0.0302		

The turnover behavior of the optimal asset policy through time is measured by the average absolute change of the asset weights. Risk is measured by the second downside moment of the funding ratio, while the costs of funding the liabilities are measured by the average contribution rate. The objective function is to minimize the costs, while adding a penalty for downside risk.

at subsequent stages is 37%. This indicates major changes in the asset mix through time; on average 1/3rd of the portfolio is turned over at each period. The average contribution rate for the asset mix recommended by this model equals 1.26%.

Given the large turnover of the optimal investment strategy, we should regard the solution with some suspicion. The sparse branching structure of the event tree is in part the culprit. At each time period there are no more than ten states to represent the underlying conditional distribution of the five time series. Moreover, these states are sampled randomly. As a result, the mean and covariance matrix will most likely be specified incorrectly in most nodes of the tree, and the optimizer chooses an investment strategy based on an erroneous approximation of the return distribution.

The random sampling procedure for constructing a sparse multi-period event tree apparently leads to unstable investment strategies. An obvious way to deal with this problem is to increase the number of nodes in the randomly sampled event tree, in order to reduce the approximation error relative to the vector autoregressive model. However, the stochastic program might become computationally intractable if we increase the number of nodes at each stage, due to the exponential growth rate of the tree. Alternatively, the switching of asset weights might be bounded by adding constraints to the model. Although we might get a more stable solution in this case, the underlying problem remains the same: the optimal decisions are based on an erroneous representation of the return distributions in the event tree.

3.4.2. Adjusted random sampling

An adjusted random sampling technique for constructing event trees can resolve some of the problems of the simple random sampling method. First, assuming an even number of nodes, we apply antithetic sampling in order to fit every odd moment of the underlying distribution. For example, if there are ten succeeding nodes at each stage then we sample five vectors of error terms from the vector autoregressive model. The error terms for the five remaining nodes are identical but with opposite signs. As a result we match every odd moment of the underlying error distributions (note that the errors have a mean of zero). Second, we rescale the sampled values in order to fit the variance. This can be achieved by multiplying the set of sampled returns for each particular asset class by an amount proportional to their distance from the mean, as in Carino et al. (1994). In this way the sampled errors are shifted away from their mean value, thus changing the variance until the target value is achieved. The adjusted values for the error terms are substituted in the estimated equations of the vector autoregressive model to generate a set of nodes for the event tree.

We now solve the model for the Dutch pension funds using the event tree generated with the adjusted random sampling method. Table 6 displays the solution. The portfolio composition in this case is 57% in stocks and 43% in bonds, and the average turnover has decreased to 12.3%. Compared to the results obtained with standard random sampling (Table 5) the average contribution rate has risen from 1.26 to 5.83%, and both downside risk and the objective value of the model have worsened.

Table 6
Solution of stochastic programming model: adjusted random sampling

Initial asset mix	Bonds	Real	Stocks
	0.00	0.43	0.57
Turnover	Bonds	Real	Stocks
	0.16	0.09	0.12
Objective value	0.5013		
Mean contribution rate	0.0583		
Downside risk	0.0375		

It seems to be more difficult to make investment profits when using an event tree constructed with adjusted random sampling, compared to plain random sampling. This is a strong signal that the profits in Table 5 were actually spurious, based on a flawed description of the underlying return distributions. A second signal is the reduction of asset mix switching due to the adjustment of the random samples. The heavy trading activity in the optimal solution of Table 5 probably results from errors in the mean returns, to which portfolio optimization models are very sensitive. Kallberg and Ziemba (1981, 1984) and Chopra and Ziemba (1993) have shown that estimation errors in the mean value can have an effect on the optimal portfolio cash-equivalent value which is ten times as large as the error introduced from mis-specification of the variances, which in turn is twice as significant as errors in the estimation of covariance matrices. Using adjusted random sampling to match the mean and the variance, we substantially reduce useless trading. The additional computational effort for adjusting the random samples is negligible.

3.4.3. Fitting the mean and the covariance matrix

A third method for constructing event trees is to estimate returns that match the first few moments of the underlying return distributions. This can be achieved by solving a non-linear optimization model following Høyland and Wallace (2001). The decision variables in the optimization model are the returns and the probabilities of the event tree, while the objective function and the constraints enforce the desired statistical properties. The probabilities and returns in all nodes of the event tree can be estimated simultaneously. However, with this approach it might take longer to construct a desirable event tree than to solve the stochastic programming model for ALM itself. The tree fitting problem can be simplified by applying the method at each stage recursively, given the assumption that the return distributions are not path dependent. This assumption is valid for long term asset and liability management with broad asset classes, but fails when modelling money management problems with path-dependent securities.

To illustrate the concepts, we write down the tree fitting equations to estimate a set of perturbations that will fit the mean, the covariance matrix, the skewness and the kurtosis of the vector autoregressive process in Table 3. The probabilities are assumed

uniform in order to ease comparison with random sampling. Let $i = 1, 2, \ldots, m$, denote the random time series that are modelled by the vector autoregressive process. In our example these are the returns on stocks, bonds, deposits, real estate and the wage growth rate. Suppose that a total of M succeeding nodes at stage $t + 1$ are available to describe the conditional distribution of these random variables in a particular node at stage t. We define the perturbation ε_{ti}^{l} as the realization in node l for the ith element of the vector ε_t, for $i = 1, 2, \ldots, m$.

We can fit the first four moments and the covariance structure of the vector autoregressive model by finding a set of perturbations that satisfy Equations (32)–(35). Equation (32) specifies that the average of the perturbations should be zero, while Equation (33) states that they should have a covariance matrix equal to Q. Additionally, Equations (34) and (35) match the skewness and the kurtosis of a normal distribution, which is the underlying return distribution of the perturbations.

$$\frac{1}{M}\sum_{l=1}^{M}\varepsilon_{ti}^{l} = 0, \quad \text{for all } i = 1, 2, \ldots, m, \tag{32}$$

$$\frac{1}{M-1}\sum_{l=1}^{M}\varepsilon_{ti}^{l}\varepsilon_{tj}^{l} = Q_{ij}, \quad \text{for all } i = 1, 2, \ldots, m,\ \ j = 1, 2, \ldots, m, \tag{33}$$

$$\frac{1}{M}\sum_{l=1}^{M}\left(\varepsilon_{ti}^{l}\right)^{3}/(Q_{ii})^{1.5} = 0, \quad \text{for all } i = 1, 2, \ldots, m, \tag{34}$$

$$\frac{1}{M}\sum_{l=1}^{M}\left(\varepsilon_{ti}^{l}\right)^{4}/(Q_{ii})^{2} = 3, \quad \text{for all } i = 1, 2, \ldots, m. \tag{35}$$

Obtaining a solution of the non-linear system (32)–(35) can be difficult, specially when higher order moments like skewness and kurtosis are included in the restrictions (see Høyland, Kaut and Wallace (2003) for a heuristic tree construction algorithm that fits the correlation matrix of a multivariate distribution and the first four moments of each marginal distribution). Instead of solving a system of nonlinear equations we may solve instead a non-linear optimization model that penalizes deviations from the desired moments in the objective function. Good starting points for this optimization can be obtained using the adjusted random sampling method of the previous subsection, which is computationally very efficient. After solving the non-linear fitting model, we can substitute the optimal set of perturbations in the estimated equations of the vector autoregressive model to generate conditional return distributions. By applying this procedure recursively, from node to node and from stage to stage, we generate an event tree that fits the time varying conditional expectation and the covariance matrix of the underlying return distributions.

We have applied this method to construct a fitted 1-10-6-6-4-4 event tree. For the later stages, with four succeeding nodes each, we fitted only the mean and covariance matrix. For the earlier stages with six or ten succeeding nodes we also matched skewness and

Table 7
Solution of stochastic programming model: fitted tree

Initial asset mix	Bonds	Real	Stocks
	0.37	0.30	0.33
Turnover	Bonds	Real	Stocks
	0.08	0.05	0.06
Objective value	0.5213		
Mean contribution rate	0.0624		
Downside risk	0.0346		

kurtosis. More nodes are used in the earlier stages, while fewer nodes in the later stages introduce less approximation error into the optimal solution of the stochastic program. Table 7 displays the optimal solution of the stochastic programming model based upon the fitted event tree. The initial asset mix now consists of 33% stocks, 30% real estate and 37% bonds. Average turnover of the investment strategy has been reduced to 6%, but the average contribution rate increased to 6.24%. Note that the optimal solutions based on adjusted random sampling and tree fitting are still quite different.

It is still difficult to conclude much about the quality of the optimal solutions, as the ALM model was solved just three times using three particular event trees. Testing of stochastic programming models requires extensive out-of-sample simulations (Dupacova, Bertocchi and Morrigia, 1998) and rolling horizon dynamic games (Golub et al., 1995). A rolling horizon simulation was used by Kouwenberg (2001) to investigate the issue of scenario generation, comparing the three different methods for constructing event trees. The results of repeated experiments for the 5-period pension fund model based on 1-10-6-6-4-4 event trees confirm that random sampling leads to poor results. Adjusted random sampling significantly improves the performance of the model. Tree fitting is only slightly better for this particular application. With event trees based on the fitting method the stochastic programming approach was shown to outperform a benchmark static model and a benchmark model with an optimized decision rule.

Finally, for completeness we would like to mention some other promising methods for constructing event trees from the stochastic programming literature. Mulvey and Zenios (1994) and Zenios (1995) discusses simulation techniques to generate scenarios of returns for fixed-income portfolio models, based on a underlying fine-grained interest rate lattice. Lattice models for scenario generation that integrate interest rate and credit risk are given in Jobst and Zenios (2005). Shtilman and Zenios (1993) derive theoretical results for the optimal sampling from lattice models. Proper spacing of the returns seems crucial for reducing approximation errors. Carino, Myers and Ziemba (1998) introduce a clustering algorithm that reduces a large random sample of asset and liability returns to a small number of representative nodes, while preserving important statistical properties such as the mean and the standard deviation of the larger sample. More recently, scenario reduction algorithms have been developed by Dupacova, Gröwe-Kuska and Römisch (2003) and Heitsch and Römisch (2003). Dupacova, Consigli and Wallace

(2000) and Pflug (2001) derive promising theoretical results for optimal scenario generation for multiperiod financial optimization. Further theoretical and empirical research in this area is important, as the event trees used as input are crucial for the effectiveness of the stochastic programming approach to ALM.

3.5. Options, bonds and arbitrage

We now turn to another requirement for scenarios used in ALM systems, the persuasive no-arbitrage condition. We review first the issues in the context of stochastic programming and then discuss the construction of arbitrage-free event trees. The generation of scenarios for ALM applications with multiple bonds or derivatives requires special attention to the no-arbitrage condition. The prices of bonds with different maturities are often driven by a small number of underlying factors such as the short term interest rate, a long term yield and the credit spread. Because of the close relationship between bond prices and interest rates, the prices movements of bonds of a similar type, but with different maturities, are often strongly associated. The price changes of derivative securities on a single underlying value, e.g., a stock index, are also often closely related.

An important concept for bond and option valuation is the construction of so called replicating strategies. For example, consider an economy with three bonds, A, B and C. Bond A matures in one year and will pay off \$110. It's current price on the market is \$105. Bond B matures in two years and will pay off \$110 too. Due to its longer maturity, bond B is cheaper: its price is \$100. A third bond, C, has a maturity of two years and will pay interest of \$11 after one year and \$110 after two years. Note that we can also create the payoff of bond C by buying $1/10$th units of bond A and 1 unit of bond B: this *replicating portfolio* has exactly the same cashflows as bond C. Hence, ideally, the price of bond C should be equal to the price of the replicating portfolio, which is $(1/10)\$105 + \$100 = \$110.5$.

If the price of bond C is not equal to \$110.5, then an *arbitrage opportunity* exists. Suppose that a market participant offers 100 units of bond C for the price of \$109.5. In the absence of transaction costs, we could make a *riskless profit* by buying 100 bonds C at the low price of \$109.5, while selling 100 units of the replicating portfolio. The sale of 100 units of the replicating portfolio delivers a payoff of \$110,500, while we only have to invest \$109,500 to buy the bonds C. After making the profit of $\$110{,}500 - \$109{,}500 = \$1000$, we never have any risk of losing money on the portfolio, as the payoffs of bond C will exactly cover the payments that have to be made for the replicating portfolio.

One of the fundamental assumptions in the theory of finance is that riskless arbitrage opportunities as described above are not available. Eventually investors will recognize these opportunities and their trading will affect prices and eliminate the riskless profits. The absence of arbitrage and the concept of replicating strategies are at the heart of modern derivative pricing methods. Absence of arbitrage opportunities is an important property for event trees of asset returns that are used as input for stochastic programming models as well. If there is an arbitrage opportunity in the event tree, then the optimal

solution of the stochastic programming model will exploit it. An arbitrage strategy creates profits without taking risk, and hence it will increase the objective value of nearly any financial planning model. It is prudent for long term ALM applications to generate scenarios that do not allow for arbitrage. If arbitrage opportunities do arise in practice then professional arbitrageurs will exploit them on a very short notice, while the focus of ALM modelling is on long term decisions.

A potential problem for stochastic programming models in ALM are arbitrage opportunities in the event tree that are due to approximation errors. Klaassen (1997) was the first to address this issue. Arbitrage opportunities might arise because the underlying return distributions are sometimes approximated poorly with a small number of nodes in the event tree. If the application only involves broad asset classes such as a stock index, a bond index and real estate index, then arbitrage opportunities are unlikely to occur unless the errors in the event tree are very big. However, applications that involve options, multiple bonds or other interest rate derivative securities can be quite vulnerable to these problems. For example, the prices of bonds with different maturities have to satisfy certain pricing relationships, as we demonstrated previously for bond A, B and C. If such a relationship is violated in a node of the event tree because of a small approximation error, then the tree contains an arbitrage opportunity and hence a source of spurious profits for a stochastic programming model.

3.5.1. Arbitrage-free event trees

Before we introduce scenario generation methods for arbitrage-free event trees we first introduce the no-arbitrage theorem from financial economics. Formally, arbitrage opportunities do not exist if there are no portfolios having a negative price that provide a non-negative payoff for all possible futures. Suppose that m primitive securities that are traded in a one period model, where the future is represented by a one period event tree with M nodes. The initial asset prices are denoted by P_{0i} and the final asset prices and payoffs are, respectively, P_{1i}^l and F_{1i}^l, for assets $i = 1, 2, \ldots, m$ and nodes $l = 1, 2, \ldots, M$ on the one period event tree. Harrison and Kreps (1979) prove that the following conditions are both necessary and sufficient for the absence of arbitrage opportunities[3]:

Theorem 1. *There are no arbitrage opportunities if and only if there exists a strictly positive probability measure $p^l > 0$, such that*

$$\sum_{l=1}^{M} p^l \frac{P_{1i}^l + F_{1i}^l}{P_{0i}} = \sum_{l=1}^{M} p^l \frac{P_{11}^l + F_{11}^l}{P_{01}}, \quad \textit{for all } i = 2, 3, \ldots, m. \tag{36}$$

[3] We assume that all initial asset prices are strictly positive. Moreover, at least one primitive asset should have a set of strictly positive prices at time 1.

If we take the base security $i = 1$ to be the one-period riskless bond, providing a continuously compounded return of r, then condition (36) reduces to:

$$P_{0i} = e^{-r} \sum_{l=1}^{M} p^l \left(P_{1i}^l + F_{1i}^l \right), \quad \textit{for all } i = 2, 3, \ldots, m. \tag{37}$$

It is clear that given a set of m primitive assets, at least $M \geqslant m$ nodes are needed in the one period event tree to satisfy the no arbitrage condition, if we would also like to avoid linearly dependent asset returns. For a multistage stochastic programming model, based on a multi period event tree, the one period no-arbitrage condition needs to hold *in each one period subtree* of the event tree. In ALM applications each one period subtree typically represents a conditional return distribution for the assets and the liabilities from time t to $t+1$. Suppose that we would like to represent the distribution of m returns in a multiperiod event tree that is used as input for a multistage stochastic programming model, then this requires at least m succeeding states in each one period subtree of the multiperiod tree. If this condition is not met, then the event tree contains a money machine for the stochastic programming model to exploit or the asset returns are linearly dependent (i.e., some of the assets are perfectly correlated).

In practice round-off errors might cause the system (37) to be infeasible, even if arbitrage profits are economically insignificant due to transaction costs. Naik (1995) provides a version of the no-arbitrage conditions that takes bid-ask spreads into account and is more easy to apply:

Theorem 2. *Suppose that assets are bought at time zero at the ask price P_{0i}^a and sold at the bid price P_{0i}^b, with $P_{0i}^a > P_{0i}^b$. Security $i = 1$ is a riskless bond, with rate r. There are no arbitrage opportunities if and only if there exists a strictly positive probability measure $p^l > 0$, such that*

$$P_{0i}^b \leqslant e^{-r} \sum_{l=1}^{M} p^l \left(P_{1i}^l + F_{1i}^l \right) \leqslant P_{0i}^a, \quad \textit{for all } i = 2, 3, \ldots, m. \tag{38}$$

Once an event tree with asset prices has been constructed, we can check for arbitrage opportunities by solving Equation (37) or (38) for each one period subtree in the multi period event tree. If we can find a set of strictly positive probabilities p^l *for each one period subtree* then the entire multiperiod event tree does not contain arbitrage opportunities. If the system happens to be infeasible for one particular subtree, then arbitrage opportunities are present and it is unwise to solve a stochastic programming model with this event tree as input.

It is possible to enforce the absence of arbitrage opportunities while constructing an event tree for asset returns. The first approach is to include the no-arbitrage condition (38) as a constraint in the tree fitting model (32)–(35), while adding the risk neutral probabilities p^l as decision variables. The probabilities are required to be strictly

positive. If we can not find a feasible solution for the tree fitting problem with these additional constraints, then arbitrage opportunities are inevitable. In this case we could try to eliminate the arbitrage opportunities by reducing the number of moment matching constraints in the tree fitting model or by increasing the number of nodes in the event tree. This entails a tradeoff between model accuracy and computational complexity.

A second approach is to start with a very fine-grained event tree of asset prices without arbitrage opportunities and then to reduce it to a smaller tree, while preserving the property of no-arbitrage. Klaassen (1998) studies a bond portfolio management problem, where the bond prices are calculated with a one-factor model for the term-structure of the interest rates. Bond prices are first generated with a binomial lattice for the one-factor interest rate model: the lattice is consistent with the initial bond prices on the market and contains no arbitrage opportunities. However, the lattice is not suited as input for a stochastic programming model, as it consists of many small time-steps in order to calculate prices of bonds and other interest rate dependent securities accurately. Each possible path on the recombining interest rate lattice from the initial period to the planning horizon would become a unique scenario in the non-recombining event tree for a stochastic programming model. As a result, the total number of scenarios in the event tree can easily get extremely huge as the number of timesteps increases.

Aggregation methods are essential to reduce a recombining lattice to a much smaller event tree with less trading dates, while preserving the property of no-arbitrage. Klaassen (1998) proposes an aggregation method that starts with a full-blown non-recombining event tree consisting of a huge number of interest rate scenarios, mirroring the information in the underlying arbitrage-free recombining binomial lattice. Recursively, the method replaces a combination of nodes at a particular time by one aggregate node, while preserving the no-arbitrage property. If a node has only one particular successor remaining at the next time, then the intermediate period can be eliminated. Eventually, the initial full-blown event tree has been reduced to a size that enables optimization of a stochastic programming model, without sacrifying the no-arbitrage property. Gondzio, Kouwenberg and Vorst (2003) propose a related aggregation method that starts with a fine-grained multi-dimensional lattice of security prices and reduces it to a sparse event tree without introducing arbitrage.

Although the absence of arbitrage opportunities is important for financial stochastic programs with derivative securities and bonds, one should keep in mind that it is only a *minimal requirement* for the event tree. The fact that the stochastic program can not generate riskless profits from arbitrage opportunities does not imply that the event tree is also a good approximation of the underlying return process. We still have to take care that the conditional return distributions of the assets and the liabilities are represented properly in each one period subtree of the multiperiod event tree. In order to avoid computational problems that arise if the tree becomes too big, one could reduce the number of stages of the stochastic program. In this way more nodes are available to describe the return distributions accurately. It is also important to include more nodes for the earlier stages, while larger errors in the later stages will have a small effect on the first-stage decisions which are the decisions implemented today by the decision makers.

End effects of stochastic programming models for ALM applications are studied by Carino and Ziemba (1998) and Carino, Myers and Ziemba (1998).

4. Comparison of stochastic programming with other methods

The literature on models for ALM is vast and dates back to the seminal contribution of Markowitz (1952). But it was not until the 1980s that the use of formal mathematical models to support financial decision making rose to wide-spread prominence in practice (Zenios, 1993). Globalization and innovations in the financial markets are the driving force behind this development that continues unabated to this date, aided by advances in computing technology and the availability of software. Four alternative modelling approaches have emerged as suitable frameworks for representing ALM problems Ziemba and Mulvey (1998). These approaches are briefly discussed in this section and compared with stochastic programming. See also the chapters by Markowitz and by Brennan and Xia in this volume for alternative ALM methodologies.

4.1. Mean-variance models and downside risk

The mean-variance framework of Markowitz (1952) is widely considered as the starting point for modern research about optimal investment. In the mean-variance framework the optimal portfolio of an investor is derived by minimizing the variance of the portfolio return, subject to a given mean return. Markowitz demonstrated that investors can reduce risk by forming well-diversified portfolios of individual stocks supporting the popular advise "Don't put all your eggs in one basket". Moreover, Markowitz (1952) makes clear that we have to pay the price of increasing risk (variance) in order to obtain a higher expected return.

The original version of the mean-variance model ignores the liability side of the investor's balance sheet. Sharpe and Tint (1990) propose an extended mean-variance model for the surplus, defined as the asset value minus the liabilities. In a surplus management model the investor minimizes the variance of the surplus return, for a given level of mean surplus return. The main result is that the optimal investment policy depends on the covariance of the asset returns with the liability return. The investor is willing to accept a lower expected return on assets that provide a higher covariance with the liability return. In order to quantify this effect, Sharpe and Tint (1990) introduce the *liability hedging credit* of an asset.

Liability hedging credits are at the core of investment strategies for asset liability management. If the investor is concerned about the net value of his balance sheet (i.e., the surplus), then he has to reckon with the correlation between asset and liability returns. Although this basic insight provided by the surplus management model of Sharpe and Tint is important, the model is of limited use in practice. There are two major drawbacks associated with mean-variance models in the context of asset and liability

management. Variance is not always a good risk measure for investors, and a single-period model might be inappropriate for multi-period investment problems with long horizons. We first turn our attention to alternative risk measures and then focus on the literature about multi-period investment models.

Intuitively it seems rather odd that variance as a risk measure penalizes positive returns and negative returns equivalently. Of course, it is only important to make a distinction between negative and positive returns if the distribution of the portfolio returns is asymmetric. However, most stock return distributions are skewed (Fama, 1965). Moreover, if we add derivatives such as call and put options to the investment opportunity set then the portfolio return distribution might also become asymmetric. In order to deal with these problems, downside-risk measures have been introduced as a substitute for variance (Bawa and Lindenberg, 1977 and Fishburn, 1977).

Downside-risk measures only penalize returns below a given threshold level, specified by the investor. Popular measures include: shortfall probability, expected shortfall, semivariance, one-sided mean absolute deviation (see Bawa and Lindenberg, 1977; Worzel, Vassiadou-Zeniou and Zenios, 1994 and Rockafellar and Uryasev, 2000). In the normative ALM literature the downside-risk concept is highly successful and seems to have replaced variance as a risk measure (Harlow, 1991; Sortino and van der Meer, 1991; Boender, 1997; Artzner et al., 1999). Investors that try to meet their liabilities usually apply a threshold of 1 for the funding ratio, i.e., the ratio of assets to liabilities. A mean-semivariance model for the surplus or the funding ratio can be solved very efficiently when formulated in an equivalent mean-absolute deviation model. This has been demonstrated by Konno and Yamazaki (1991) and Konno and Kobayashi (1997).

4.2. Discrete-time multi-period models

Apart from the criticism of the risk measure in the mean-variance model, a second concern is that the model only considers a single period, without opportunities to change the investment strategy inter-temporally. Most asset liability management problems cover a long period of time until the planning horizon, with multiple opportunities to change the investment portfolio. For example, the planning horizon for most insurance products extends beyond a decade, for pension funds it is more than 30 years, and for social security plans it may go up to 50 years. It is clearly inappropriate to model several decades into the future in one single period without allowing for trading to adjust the investment portfolio.

Several authors have studied models that overcome the single-period restriction of the mean-variance model. Mossin (1968), Samuelson (1969) and Hakansson (1970) study the multi-period consumption-investment problem for investors maximizing the expectation of a power utility function over wealth at the end of their horizon, under the assumption of intertemporally independent distributed asset returns. A power utility function has the property of constant relative risk aversion and as a result the portfolio weights are equal in each period, regardless of the individual's age and wealth. The

investment policies are called myopic as the investor behaves identically to a single-period investor. Mossin (1968) showed that a sequence of myopic models are optimal for a multi-period investor under the assumptions of intertemporal independence of returns, no transaction costs and no cash infusion or withdrawals. Hakansson (1969, 1971) generalizes Mossin's results for logarithmic utility functions and adds stochastic wage income, serially correlated asset returns and uncertainty about the individual's lifetime to the basic setup.

The multi-period investment models of Mossin (1968), Samuelson (1969) and Hakansson (1970) are solved in closed-form with dynamic programming. Dynamic programming for ALM can lead to interesting insights, as the optimal policies are derived in feed-back form. However, if we would like to consider practical issues like transaction costs, trading restrictions and non-normal return distributions, then closed-form solution are difficult to obtain. A numerical implementation of the dynamic programming algorithm can be applied to solve ALM problems (see Berkelaar and Kouwenberg, 2003), however the curse of dimensionality severely limits the type of models that can be solved in a reasonable amount of time. Models with more than three or four state variables are bound to run into serious computational problems, and it is difficult to handle transaction costs.

An alternative approach is to specify a *decision rule* for changing the investment strategy in advance, and optimize the parameters of the given rule. For instance, Maranas et al. (1997) use this approach for a multi-period fixed mix model, which optimizes a set of fixed asset weights to maximize the objective function of a multi-period ALM model. The ALM simulation system of Boender (1997) optimizes the parameters of decision rules that update the strategic asset mix and the contribution rate as a function of the current funding ratio. Models with decision rules, such as fixed mix, have several advantages. First, they can handle transaction costs and operational, regulatory or corporate restrictions on the investment policy. Moreover the optimized rule can be easily interpreted and understood by decision makers.

A major problem of models with decision rules is that the objective function is often non-unimodal. Multiple local optimal solutions may exist, and global optimization algorithms have to be applied to find the overall optimum solution. Recent progress on global optimization nonwithstanding these algorithms can often handle only a small number of decision variables (i.e., coefficients of the decision rules). Finally, we are never sure that a given decision rule is actually optimal for the problem at hand. For example, we do not know in advance that a fixed mix rule is optimal for a particular ALM problem. It might as well be that another dynamic investment rule is more efficient: for example, we could adjust the asset mix as a function of the ratio of assets to liabilities. If we apply the latter rule, there is still an immense number of different functional forms that we could choose, as the optimal relationship between the asset weights and the funding ratio can be non-linear. A disadvantage is therefore that we might have to try a large number of different specifications for the decision rule, before we are confident that we have found a rule with relatively good performance.

4.3. Continuous-time models

Continuous-time models play an important role in modern finance (Merton, 1990). As a companion paper of Samuelson (1969), Merton (1969) formulates the consumption-investment problem in a continuous-time framework, where the timestep between consecutive trading dates decreases to zero in the limit. The asset prices are assumed to follow geometric Brownian motions, which corresponds to a log-normal return distribution. For the class of power utility functions Merton (1969) demonstrates that myopic investment policies are optimal, analogous to the findings of Samuelson (1969). In a more general consumption-investment setup Merton (1971) confirms the findings of Hakansson (1969, 1971). An interesting result of Merton (1969, 1971) is *fund-separation*. The optimal portfolio of every investor can be separated into a small number of mutual funds: the riskless asset, the growth-optimal portfolio of risky assets, and a hedge-portfolio for each external source of risk that affects the asset returns or the utility of the investor.

The continuous-time models of Merton (1969, 1971) are of limited practical value for institutional asset and liability management. The assumptions about the utility function and asset prices are restrictive. More importantly in the context of ALM for large institutions this model ignores transaction costs and does not incorporate trading restrictions that may be imposed by regulators or dictated by corporate policy. The financial economics literature has lately paid more attention to previously ignored "details" that are very relevant for practitioners, resorting to numerical techniques if necessary. For example, Brennan, Schwartz and Lagnado (1997) numerically investigate the impact of return predictability on optimal portfolio choice in a continuous-time investment model based on Merton (1969). The optimal portfolio weights reported by Brennan, Schwartz and Lagnado (1997) tend to fluctuate drastically through time, resembling "yoyo-strategies", due to the absence of transaction costs in the model and the lack of uncertainty about the model parameters.

Balduzzi and Lynch (1999, 2000) show that transaction costs can indeed stabilize the optimal policy of an optimal control model with return predictability. In practice transaction costs might not be the only concern, but also uncertainty about the actual value of model parameters such as the mean asset return. Brennan (1998), Barberis (2000) and Xia (2000) study continuous-time investment models under parameter uncertainty, with dynamic learning about parameter values (Brennan, 1998), with return predictability (Barberis, 2000) and with both learning and predictability (Xia, 2000). Another recent development is that optimal portfolio and consumption problems with return predictability can be solved in closed-form, as demonstrated by Kim and Omberg (1996) and Liu (1999). See also the chapter by Brennan and Xia in this volume.

An advantage of the continuous-time framework is that optimal decision rules can be derived for some basic models: Cairns and Parker (1997) and Rudolf and Ziemba (2004) derive optimal decision rules in closed form for small ALM problems. Moreover, the impact of transaction costs, trading limits, return predictability, parameter uncertainty and market incompleteness can be analyzed quite accurately in models that focus on

one or two of these issues in isolation. However, general models that incorporate all of these issues simultaneously have not been solved yet. Similarly, practical constraints reflecting regulatory restrictions, operational requirements or corporate policy have not been incorporated. Moreover, an attempt to solve such a general model is very likely to run into computational problems due to the curse of dimensionality.

4.4. Stochastic programming

Continuous-time models and discrete-time models solved with dynamic programming and optimal control can provide good qualitative insights about fundamental issues in investments and ALM, as the optimal decision rules are in feed-back form. However, their practical use as a tool for decision making is limited by the many simplifying assumptions that are needed to derive the solutions in a reasonable amount of time. The stochastic programming approach for ALM discussed in this chapter can be considered as a practical multi-period extension of the normative investment approach of Markowitz (1952). The advantage of stochastic programming models for multi-period investment and ALM problems is that important practical issues such as transaction costs, multiple state variables, market incompleteness, taxes and trading limits, regulatory restrictions and corporate policy requirements can be handled *simultaneously* within the framework.

Of course this flexibility comes at a price and stochastic programming also has a drawback. The computational effort explodes as the number of decision stages in a multi-stage stochastic programming model increases. While implementing a stochastic programming model for ALM, we are therefore often forced to make a trade off between the number of decision stages in the model and the number of nodes in the event tree that are used to approximate the underlying returns distributions. While setting up stochastic programming models it is important to keep in mind that a normative model does not necessarily have to include every possible decision moment up to the planning horizon. Capturing the first few opportunities accurately can be good enough to make an informed decision right now. End effects created by limiting the number of stages can be mitigated using the techniques discussed in Carino and Ziemba (1998) and Carino, Myers and Ziemba (1998).

It is not an exaggeration to claim that stochastic programming can deal simultaneously with all important aspects of an ALM system. However, even if the model would be solvable, too many details would confuse instead of support the decision maker. Like the alternative methodologies discussed in this section, stochastic programming applications to ALM have a strong element of art. Stochastic programming seems to enjoy several advantages over the alternatives but it is not without shortcomings.

5. Applications of stochastic programming to ALM

In recent years the number of publications about stochastic programming for asset liability management has risen drastically, probably inspired by the radical increase of

efficiency and accessibility of computer systems. Ziemba and Mulvey (1998) categorize the models in three generations: (i) model origins, that deal with the early mathematical formulations, (ii) early models, that deal with real-world applications but developed and tested in a limited setting, mostly by academic researchers, and (iii) modern models, that deal with a variety of institutional problems, developed as large-scale applications and tested extensively, usually in collaboration with institutional asset and liability managers. From the early models we mention the bond portfolio management model of Bradley and Crane (1972), the bank ALM model of Kusy and Ziemba (1986), the fixed income model of Zenios (1991), the asset allocation model of Mulvey and Vladimirou (1992) and the stochastic dedication model of Hiller and Eckstein (1993).

Under modern models Ziemba and Mulvey list around forty references. We mention a sample that covers a broad range of applications, selecting publications where the commercial component of the model was substantial. The insurance ALM model of Carino and co-authors (1994, 1998) has been used extensively by the Frank Rusell company in consulting ALM managers in insurance and pension fund. Their work with the The Yasuda Fire and Marine Insurance Company (Japan) was a finalist at the Franz Edelman Competition for Management Science Achievements. Similar acclaim was achieved by the Towers Perrin–Tillinghast model of Mulvey, Gould and Morgan (2000). Stochastic programming models for Dutch pension funds were developed by Dert (1995), a general ALM model for insurers by Consigli and Dempster (1998) and an application to the Norwegian insurance industry by Høyland and Wallace (this volume). Models for money management with mortgage backed securities were developed by Golub et al. (1995) and for insurance products by Nielsen and Zenios (1996a, 1996b). A multi-period model, but without portfolio rebalancing decisions, for insurance products with minimum guarantee is discussed in the chapter by Consiglio, Cocco and Zenios in this volume.

The relevance of a normative approach can only be judged on the basis of actual performance. Carino, Myers and Ziemba (1998) report good performance of the Frank Rusell ALM model applied to Yasuda Insurance, with yearly savings of up to $79 million compared to a simple constant mix strategy. Mulvey, Gould and Morgan (2000) report an estimate of $450 to $1000 million savings in opportunity costs using the Towers Perrin–Tillinghast model to plan the US West pension plan. Golub et al. (1995), Kouwenberg (2001) and Fleten, Høyland and Wallace (2002) apply simulation in order to test the performance of stochastic programming models, relative to simple one-period models. The results indicate that dynamic stochastic programming models can outperform simple mean-variance models and multi-period models with fixed mix decision rules. Significant savings on transaction costs achieved with stochastic programming models were also demonstrated by Mulvey (1993). The effectiveness of stochastic programming models for tracking broad market indices has been demonstrated by Consiglio and Zenios (2001), Zenios et al. (1998) and Worzel, Vassiadou-Zeniou and Zenios (1994).

Golub et al. (1995) apply both out-of-sample simulations and dynamic, rolling horizon games, to compare the performance of stochastic programming models against

single-period myopic models. This is, to the best of our knowledge, the first study to compare these two classes of models using out-of-sample simulations and dynamic games. In particular they tested a mean-absolute deviation model against stochastic programming for money management problems with mortgage-backed securities. The objective of the money manager was to achieve a target return in excess of the 3-year Treasury bond benchmark by investing in a diversified portfolio of mortgage backed securities and government bonds. Uncertainty in interest rates and mortgage prepayments are the main risk factors in this setting. In both out-of-sample simulations and dynamic games a two-stage stochastic programming model consistently outperformed the single period model. The stochastic programming model was also found to be more robust than mean-absolute deviation with respect to out-of-sample changes in volatility. Golub et al. also tested the performance of the model against the popular fixed-income portfolio immunization technique (Reddington, 1952; Christensen and Fabozzi, 1987; Zenios, 1993). Their findings confirm the conclusions of Mulvey and Zenios (1994) that modern fixed income securities are best managed by capturing correlations—in a single-period setting in this study—instead of using duration matching in an immunization framework.

The work of Golub et al. was extended by Zenios et al. (1998) to multistage models. Their empirical investigations were once more carried out using out-of-sample simulations and rolling horizon dynamic games, but this time comparing single period with two-stage and three-stage models. The three-stage model outperformed the two-stage model for 12 out of the 15 repetitions of the dynamic game, over a three year period. Summary statistics are reported in Table 8.

This model was also tested, *ex post* in tracking the Salmon Brothers index of mortgage-backed securities in the US. Backtesting was carried out over the three year period January 1989 to December 1991, in monthly steps. During this period the index realized an annualized return of 14.05%, the single-period model (mean-absolute deviation) 14.18% and the two-stage stochastic program 15.10%. Returns of the model

Table 8
Zenios et al. (1998): dynamic games with stochastic programming

Statistic	Single-period model	Two-stage stoch. prog.
Min. return	8.11	7.74
Max. return	11.61	9.78
Mean return	8.60	8.83
Std. dev.	0.86	0.59

The table shows the results of dynamic rolling horizon games for money management with mortgage backed securities. The target is to exceed the benchmark return of the 3-year Treasury and achieve highest expected return during the holding period. A three-year horizon is considered and the experiments were repeated 15 times with a three-year rolling horizon and portfolio rebalancing every six months.

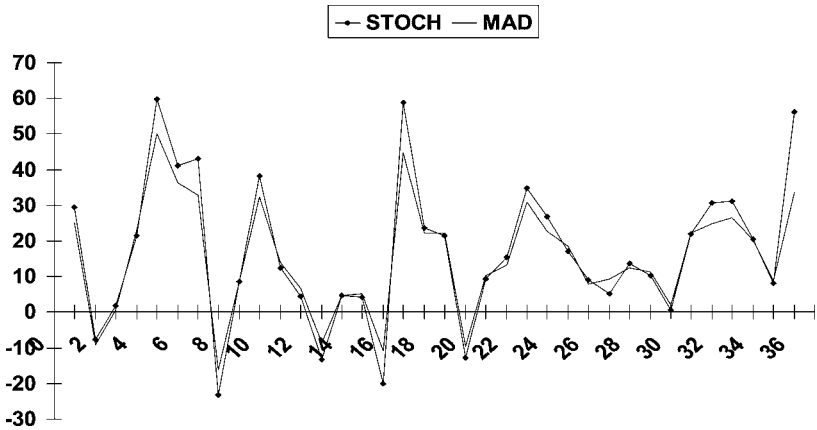


Fig. 4. Tracking error of a singe-period and a stochastic programming model in tracking the Salomon Brothers index of mortgage backed securities during the 36-month period January 1989 to December 1991.

portfolios account for transaction costs. Figure 4 illustrates the performance of both models.

Gondzio, Kouwenberg and Vorst (2003) apply a stochastic programming model to an option hedging problem in an incomplete market with stochastic volatility and transaction costs. They test the performance of the stochastic programming model and other hedging strategies with simulations. The test problem assumes that an investor has sold a call option on the stock index with a maturity of one year, for a price of $8.41. This liability is unique: it can not be traded on an organized market. The aim of the investor is to minimize the expected negative hedging errors, which occur if the value of the hedging portfolio drops below the value of the liability. In order to hedge his liability the investor can trade stocks, borrow and lend money, and trade liquid short-term call and put options on the index with a maturity of 3 months. All trades lead to transactions costs due to differences in bid and ask prices: 0.25% for trading the index (futures), 2.5% for trading short-term options and there is a spread of 0.5% between borrowing and lending.

The stochastic programming model for the hedging problem consists of three trading dates, covering the first three weeks of the investor's problem. The model does not necessarily have to incorporate every trading date up to the maturity of the liability, as it can be solved again each period (assuming end effects are modelled properly, see also Carino et al., 1994, 1998). Gondzio, Kouwenberg and Vorst (2003) compare the optimal strategy of the stochastic programming model to a delta hedging strategy and a delta-vega hedging strategy, which are decision rules that are traditionally applied (see Hull, 2000). The delta hedging strategy eliminates exposure to small movements of the stock price, but it ignores the effects of stochastic volatility. As the value of the liability is very sensitive to changes in the volatility of the stock index, this strategy can lead to large hedging errors. The delta-vega hedging strategy eliminates this volatility

exposure by taking a position in exchange-traded short-term options. Note that both strategies require frequent adjustments to keep the hedge up to date and this could lead to considerable transaction costs.

Gondzio, Kouwenberg and Vorst (2003) sample a total of 100,000 simulation paths for the stock index and its volatility to compare these two hedging strategies with the optimal strategy of a stochastic programming model. Panel A of Table 9 shows the simulation results of the delta hedging strategy, panel B shows the performance of the delta-vega hedging strategy and panel C presents the stochastic programming approach. The results clearly show that the delta hedging strategy (panel A) performs very badly: its average negative hedging error is approximately 100 times worse than the hedging error of the stochastic programming model. This poor performance comes as no surprise, given that a pure delta hedging strategy ignores volatility movements. The delta-vega hedging strategy in panel B performs much better, because it leads to lower hedging errors and it entails less trading. The stochastic programming approach in panel C improves the results even further by reducing transaction costs. The optimal strategy of the

Table 9
Gondzio, Kouwenberg and Vorst (2003): simulation results

Date	T-costs	Turn-over (%)	Delta gap (%)	Vega gap (%)
A. Delta hedging: hedging error 1.609%				
1	0.1418	688	0	100
2	0.0154	83	0	100
3	0.0153	84	0	100
B. Delta-vega hedging: hedging error 0.075%				
1	0.1284	292	0	0
2	0.0125	27	0	0
3	0.0273	48	0	
C. Stochastic program: hedging error 0.017%				
1	0.1354	247	0.24	0.70
2	0.0070	24	0.87	5.59
3	0.0034	12	1.51	10.35

The table shows the results of simulations with different hedging strategies. Panel A represents delta-hedging with stock trading only. Panel B represents delta-vega-hedging with a short-term at-the-money option and the stock. Panel C represents the optimal trading strategy of the SOH model, involving the five available traded short-term options and the stock (see Gondzio, Kouwenberg and Vorst, 2003 for details). The hedging error denotes the average negative hedging error after three weeks, denoted as a percentage of the initial value of the liability ($8.41). Next, for the three trading date are shown: the average transaction costs in cents, the average portfolio turn-over, the average absolute delta gap and the average absolute vega gap. The portfolio turn-over is defined as the sum of the absolute changes of the money invested in each asset divided by two times the value of the portfolio. The absolute delta (vega) gap is the absolute difference between the delta (vega) of the hedge portfolio and the delta (vega) of the 1-year call option.

stochastic program is on average quite close to a delta-vega hedging strategy, but with additional slack in order to avoid needless trading costs. In practice portfolio managers and traders might additionally face limits on the amounts they can borrow and sell short. These restrictions can be easily incorporated in a stochastic programming model, which is another advantage of the approach.

6. Solution methods and computations

Stochastic programming models grow in size very quickly with the number of stages and the number of scenarios at each stage. Some of the problems solved in the literature have equivalent deterministic formulations with hundreds of thousands of variables and constraints. For instance, the model used in previous sections to test different event trees has an equivalent linear programming formulation with 24,614 constraints, 32,100 variables and 96,586 (0.012%) non-zeros in the constraint matrix. Due to the exponential growth of the number of nodes in the event tree, the number of variables and constraints is huge. The deterministic equivalent linear program of the stochastic programming model is solved using an interior point algorithm that exploits the sparse block-angular structure. The fixed income models and the asset allocation models can also be represented as network flow problems and can be solved using special purpose network optimization algorithms (Mulvey and Vladimirou, 1992; Nielsen and Zenios, 1996a, 1996b). Table 10 summarizes the characteristics of some of the models we have cited.

Problems of this size need some specialized solution algorithms. The state-of-the-art in optimization software allows the solution of large-scale problems although the computer resources required may be substantial. In general one should expect several hours of computer time on a dedicated high-performance workstation. Almost real-time solutions have been reported in the literature for these problems, but such performance was invariably made possible with the use of parallel computers.

Table 10
Large scale ALM applications of stochastic programming

Application	Stages	Scenarios	Deterministic equivalent	Source
Fixed income	3	16,392	492,036 × 1,377,676	Nielsen and Zenios (1996a, 1996b)
Fixed income	5	6461	51,082 × 154,178	Nielsen and Zenios (1996a, 1996b)
Insurance	5	1024	123,425 × 174,480	Cariño et al. (1994, 1998)
Insurance	10	2688	352,114 × 671,961	Consigli and Dempster (1998)
Pension funds	5	5760	24,614 × 32,100	Kouwenberg (2001)
Pension funds	6	4,826,809	12,469,250 × 24,938,502	Gondzio and Kouwenberg (2001)

The special block-angular structure of the constraint matrix of stochastic programs has prompted the development of specialized algorithms. Modern implementations of the simplex method, such as IBM's OSL or CPLEX by Ilog, incorporate many theoretical results of research on this problem, making commercially available two excellent versions of this algorithm for the solution of stochastic programming problems. Similarly, interior point methods have been specialized for block angular structures and both OSL and CPLEX implement this feature as well.

Finally we mention the development of special-purpose decomposition algorithms for breaking up the deterministic equivalent formulation into smaller problems. These can be solved either serially or in parallel. In any event they are much smaller than the original problem hence solution times are substantially improved. OSL supports some decomposition methods. However, most software implementations of decomposition methods are supported by academic researchers. In general such systems are very efficient and quite robust, but they are not of industrial quality.

Work on the solution of stochastic programs has also focused on the intelligent sampling and pruning of the event tree. Clearly not all events on an event tree will have an effect on the optimal solution. It is important to sample only those events that have the most impact on the solution. Importance sampling (Dantzig and Infanger, 1991) and EVPI (expected value of perfect information, Dempster and Gassmann, 1991) have appeared as promising avenues for restricting the tree size, and structuring problems of moderate size. For a discussion of solution techniques and an extensive list of references see Censor and Zenios (1997, Chapter 13).

7. Summary and open issues

There are still many interesting open issues in the area of stochastic programming and asset liability management.

Perhaps the most important issue is in expanding the applicability of stochastic programming to address enterprise-wide risk management problems. The first step in this direction is to shape firm-wide risk analysis according to a portfolio approach. What risks arise in the operation of the business? What risks are connected to the core activities of the firm (core exposures) and what to the facilitating activities (peripheral exposures)? How do these separate risks interact on the firm-wide level and how can their aggregate influence on the business performance be analyzed? What is the contribution of these risks to the firm's (diversified) overall risk profile? Once these questions are addressed—and these are mostly pricing and simulation questions—we need to design firm-wide risk management recognizing portfolio diversification benefits and natural hedges. Stochastic programming is the ideal tool for synthesizing the firm-wide risk analysis into firm-wide risk management. Finally post-optimality analysis of the models will allow us to develop of a firm-wide performance measurement system and allowing for the decomposition of risk measures and performance measures into components attributable to the various underlying risk factors.

With regard to the generation of scenarios, an important issue is how to measure the approximation error of the returns in the event tree compared to the true underlying distribution. Once appropriate measures have been identified, one could try to develop methods for constructing event trees that minimize the approximation error (assuming the size of the event tree is fixed). Promising research in this area includes Pflug (2001), Dupacova, Gröwe-Kuska and Römisch (2003) and Heitsch and Römisch (2003). Postoptimality analysis Dupacova, Bertocchi and Morrigia (1998) also holds great potential in this respect.

With regard to the computational side of stochastic programming, there seems to be a need for flexible and efficient model generation tools. Specialized optimization algorithms and the ever increasing computational power of computers make it feasible to solve large scale multi-stage ALM models with millions of variables and constraints on desktop computers nowadays. However, most commercial mathematical modelling languages are not capable of generating the data of these huge problems efficiently. Moreover, if the modelling language does not exploit the special structure of the stochastic program, it can easily run in to memory problems that could be avoided. Model generation seems to have become the bottleneck that limits the size of multi-stage stochastic programming models applied to ALM.

Finally, with regard to designing models for actual ALM problems, more research into end-effects might be helpful. Many ALM problems in practice are long term in nature and have much more decision moments than can be captured in a single multi-stage stochastic programming model. If we only consider the first few decision moments in a stochastic programming model, it is very important to choose an objective that is consistent with the long run goals of the company and makes sure that business can also continue as usual after the planning horizon of the model. Carino et al. (1994) and Carino, Myers and Ziemba (1998) make some promising contributions in this direction.

Appendix A. Basics of probability spaces

We give some basic definitions needed in this chapter. Additional background material can be found in Billingsley (1995) and, with emphasis on stochastic programming in Wets (1989). Boldface Greek characters denote random vectors.

Let Ω be an arbitrary space or set of points. A σ-*field* for Ω is a family Σ of subsets of Ω such that Ω itself, the complement with respect to Ω of any set in Σ, and any union of countably many sets in Σ are all in Σ. The members of Σ are called *measurable* sets or *events*. The set Ω with the σ-field Σ is called a *measurable space* and is denoted by (Ω, Σ).

Let Ω be a (linear) vector space and Σ a σ-field. A *probability measure* P on (Ω, Σ) is a real-valued function defined over the family Σ, which satisfies the following conditions: (i) $0 \leqslant P(A) \leqslant 1$ for $A \in \Sigma$; (ii) $P(\emptyset) = 0$ and $P(\Omega) = 1$; and (iii) if $\{A_k\}_{k=1}^{\infty}$ is a sequence of disjoint sets $A_k \in \Sigma$ and if $\bigcup_{k=1}^{\infty} A_k \in \Sigma$ then $P(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} P(A_k)$. The triplet (Ω, Σ, P) is called a *probability space*. The

support of (Ω, Σ, P) is the smallest subset of Ω with probability 1. If the support is a countable set then the probability measure is said to be discrete. The term *scenario* is used for the elements of Ω of a probability space with a discrete distribution.

A proposition is said to hold *almost surely* if it holds on a subset $A \subseteq \Omega$ with $P(A) = 1$. The *expected value* of a random variable $\mathcal{Q}$ on (Ω, Σ, P) is the Stieltjes integral of $\mathcal{Q}$ with respect to the measure P:

$$E[\mathcal{Q}] \doteq \int \mathcal{Q}\,\mathrm{d}P = \int_{\Omega} \mathcal{Q}(\boldsymbol{\omega})\,\mathrm{d}P(\boldsymbol{\omega}).$$

Let (Ω, Σ, P) be a probability space and suppose that $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_K$ is a finite partition of the set Ω. From this partition we form a σ-field $\mathcal{A}$ which is a subfield of Σ. The *conditional expectation* of the random variable $\mathcal{Q}(\boldsymbol{\omega})$ on (Ω, Σ, P) given $\mathcal{A}$ at $\boldsymbol{\omega}$ is denoted by $E[\mathcal{Q}|\mathcal{A}]$ and defined as

$$E[\mathcal{Q}|\mathcal{A}] \doteq \frac{1}{P(\mathcal{A}_i)} \int_{\mathcal{A}_i} Q(\boldsymbol{\omega})\,\mathrm{d}P(\boldsymbol{\omega})$$

for $\boldsymbol{\omega} \in A_i$, assuming that $P(\mathcal{A}_i) > 0$.

References

Artzner, P., Delbaen, F., Eber, J.-M., Heath, D., 1999. Coherent measures of risk. Mathematical Finance 9, 203–228.

Asay, M.R., Bouyoucos, P.J., Marciano, A.M., 1993. An economic approach to valuation of single premium deferred annuities. In: Zenios, S.A. (Ed.), Financial Optimization. Cambridge University Press, Cambridge, UK, pp. 100–135.

Balduzzi, P., Lynch, A., 1999. Transaction costs and predictability: Some utility cost calculations. Journal of Financial Economics 52, 47–78.

Balduzzi, P., Lynch, A., 2000. Predictability and transaction costs: The impact on rebalancing rules and behavior. Journal of Finance 55, 2285–2309.

Barberis, N., 2000. Investing for the long run when returns are predictable. Journal of Finance 55, 225–264.

Bawa, V.S., Lindenberg, E.B., 1977. Capital market equilibrium in a mean-lower partial moment framework. Journal of Financial Economics 5, 189–200.

Beale, E.M.L., 1955. On minimizing a convex function subject to linear inequalities. Journal of the Royal Statistical Society 17, 173–184.

Berkelaar, A., Kouwenberg, R., 2003. Retirement saving with contribution payments and labor income as a benchmark for investments. Journal of Economic Dynamics and Control 27, 1069–1097.

Billingsley, P., 1995. Probability and Measure. John Wiley and Sons, New York.

Birge, J.R., Louveaux, F., 1997. Introduction to Stochastic Programming. Springer-Verlag, Heidelberg.

Black, F., Derman, E., Toy, W., 1990. A one-factor model of interest rates and its application to treasury bond options. Financial Analysts Journal, 33–39, January/February.

Boender, G.C.E., 1997. A hybrid simulation/optimisation scenario model for asset/liability management. European Journal of Operational Research 99, 126–135.

Bradley, S.P., Crane, D.B., 1972. A dynamic model for bond portfolio management. Management Science 19, 139–151, October.

Brennan, M.J., 1998. The role of learning in dynamic portfolio decisions. European Economic Review 1, 295–306.

Brennan, M.J., Schwartz, E.S., 1982. An equilibrium model of bond pricing and a test of the market efficiency. Journal of Financial and Quantitative Analysis XVII, 301–329, September.

Brennan, M.J., Schwartz, E.S., Lagnado, R., 1997. Strategic asset allocation. Journal of Economic Dynamics and Control 21, 1377–1403.

Cagan, L.D., Carriero, N.S., Zenios, S.A., 1993. A computer network approach to pricing mortgage-backed securities. Financial Analysts Journal, 55–62, March/April.

Cairns, A.J.G., Parker, G., 1997. Stochastic pension fund modelling. Insurance: Mathematics and Economics 21, 43–79.

Carino, D.R., Ziemba, W.T., 1998. Formulation of the Russell–Yasuda Kasai financial planning model. Operations Research 46, 433–449.

Carino, D.R., Myers, D.H., Ziemba, W.T., 1998. Concepts, technical issues, and uses of the Russell–Yasuda Kasai financial planning model. Operations Research 46, 450–462.

Carino, D.R., Kent, T., Myers, D.H., Stacy, C., Sylvanus, M., Turner, A.L., Watanabe, K., Ziemba, W.T., 1994. The Russell–Yasuda Kasai model: An asset/liability model for a Japanese insurance company using multistage stochastic programming. Interfaces 24 (1), 29–49.

Censor, Y., Zenios, S.A., 1997. Parallel Optimization: Theory, Algorithms, and Applications. Series on Numerical Mathematics and Scientific Computation. Oxford University Press, New York.

Chopra, V.K., Ziemba, W.T., 1993. The effect of errors in means, variances and covariances on optimal portfolio choice. Journal of Portfolio Management, 6–11, Winter.

Christensen, P.E., Fabozzi, F.J., 1987. Bond immunization: An asset liability optimization strategy. In: Fabozzi, F.J., Pollack, I.M. (Eds.), The Handbook of Fixed Income Securities. Dow Jones Irwin.

Consigli, G., Dempster, M.A.H., 1998. The CALM stochastic programming model for dynamic asset and liability management. In: Ziemba, W.T., Mulvey, J.M. (Eds.), Worldwide Asset and Liability Modeling. Cambridge University Press, Cambridge, UK, pp. 464–500.

Consiglio, A., Zenios, S.A., 2001. Integrated simulation and optimization models for tracking international fixed income indices. Mathematical Programming 89 (2), 311–339.

Consiglio, A., Cocco, F., Zenios, S.A., 2001. The value of integrative risk management for insurance products with guarantees. Journal of Risk Finance, 1–11, Spring.

Dantzig, G.B., 1955. Linear programming under uncertainty. Management Science 1, 197–206.

Dantzig, G.B., Infanger, G., 1991. Large-scale stochastic linear programs: Importance sampling and Benders decomposition. Report sol 91-4, Department of Operations Research, Stanford University.

Dempster, M.A.H., Gassmann, H.I., 1991. Stochastic programming: Using the expected value of perfect information to simplify the decision tree. In: Proceedings of 15th IFIP Conference on System Modelling and Optimization, Zurich, pp. 301–303.

Dert, C., 1995. Asset liability management for pension funds. PhD thesis, Erasmus University, Rotterdam, Netherlands.

Duffie, D., Singleton, K., 1993. Simulated moments estimation of Markov models of asset prices. Econometrica 61, 929–952.

Dupacova, J., Bertocchi, M., Morrigia, V., 1998. Postoptimality of scenario based financial planning models with an application to bond portfolio management. In: Ziemba, W.T., Mulvey, J.M. (Eds.), World Wide Asset and Liability Management. Cambridge University Press, Cambridge, UK, pp. 263–285.

Dupacova, J., Consigli, G., Wallace, S.W., 2000. Scenarios for multistage stochastic programs. Annals of Operations Research 100, 25–53.

Dupacova, J., Gröwe-Kuska, N., Römisch, W., 2003. Scenario reduction in stochastic programming: An approach using probability metrics. Mathematical Programming 95, 493–511.

Embrechts, P., 2000. Actuarial versus financial pricing of insurance. The Journal of Risk Finance 1 (4), 17–26.

Fama, E.F., 1965. The behavior of stock-market prices. Journal of Business 38, 34–105.

Fishburn, P.C., 1977. Mean-variance risk analysis with risk associated with below target returns. American Economic Review 67, 116–126.

Fleten, S.-E., Høyland, K., Wallace, S.W., 2002. The performance of stochastic dynamic and fixed mix portfolio models. European Journal of Operational Research 140, 37–49.

Goldstein, A.B., Markowitz, B.G., 1982. Sofasim: A dynamic insurance model with investment structure, policy benefits and taxes. Journal of Finance 37, 595–604.
Golub, B., Holmer, M., McKendall, R., Pohlman, L., Zenios, S.A., 1995. Stochastic programming models for money management. European Journal of Operational Research 85, 282–296.
Gondzio, J., Kouwenberg, R., 2001. High-performance computing for asset liability management. Operations Research 49, 879–891.
Gondzio, J., Kouwenberg, R., Vorst, A.C.F., 2003. Hedging options under transaction costs and stochastic volatility. Journal of Economic Dynamics and Control 27, 1045–1068.
Green, W.H., 1990. Econometric Analysis. MacMillan, New York.
Hakansson, N.H., 1969. Optimal investment and consumption strategies under risk, an uncertain lifetime and insurance. International Economic Review 10, 443–466.
Hakansson, N.H., 1970. Optimal investment and consumption strategies under risk for a class of utility functions. Econometrica 38, 587–607.
Hakansson, N.H., 1971. On optimal myopic portfolio policies, with and without serial correlation of yields. Journal of Business 44, 324–334.
Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press, Princeton, NJ.
Hansen, L., Scheinkman, J., 1995. Back to the future: Generating moment implications for continuous-time Markov processes. Econometrica 63, 767–804.
Harlow, W., 1991. Asset allocation in a downside risk framework. Financial Analyst Journal 47, 28–40.
Harrison, J.M., Kreps, D.M., 1979. Martingales and arbitrage in multiperiod securities markets. Journal of Economic Theory 20, 381–408.
Heitsch, H., Römisch, W., 2003. Scenario reduction algorithms in stochastic programming. Computational Optimization and Applications 24, 187–206.
Hiller, R.S., Eckstein, J., 1993. Stochastic dedication: Designing fixed income portfolios using massively parallel Benders decomposition. Management Science 39 (11), 1422–1438.
Holmer, M.R., 1994. The asset/liability management system at Fannie Mae. Interfaces 24 (3), 3–21.
Holmer, M.R., Zenios, S.A., 1995. The productivity of financial intermediation and the technology of financial product management. Operations Research 43 (6), 970–982.
Høyland, K., 1998. Asset liability management for a life insurance company: A stochastic programming approach. PhD thesis, Norwegian University of Science and Technology, Trondheim, Norway.
Høyland, K., Wallace, S.W., 2001. Generating scenario trees for multi-stage decision problems. Management Science 47, 295–307.
Høyland, K., Kaut, M., Wallace, S.W., 2003. A heuristic for moment-matching scenario generation. Computational Optimization and Applications 24, 169–185.
Hull, J., 2000. Options, Futures and Other Derivatives, fourth ed. Prentice-Hall.
Hull, J., White, A., 1990. Pricing interest rate derivative securities. Review of Financial Studies 3, 573–592.
Jamshidian, F., Zhu, Y., 1997. Scenario simulation: Theory and methodology. Finance and Stochastics, 43–67.
Jobst, N.J., Zenios, S.A., 2005. On the simulation of interest rate and credit risk sensitive securities. European Journal of Operational Research 161, 298–324.
Judge, G.G., Hill, R.C., Griffiths, W.E., Lütkepohl, U., Lee, T.C., 1988. The Theory and Practice of Econometrics. John Wiley and Sons, New York.
Kall, P., Wallace, S.W., 1994. Stochastic Programming. John Wiley & Sons, New York.
Kallberg, J.G., Ziemba, W.T., 1981. Remarks on optimal portfolio selection. In: Bamberg, G., Opitz, O. (Eds.), In: Methods of Operations Research, vol. 44. Oelgeschlager, Gunn and Hain, pp. 507–520.
Kallberg, J.G., Ziemba, W.T., 1984. Mis-specification in portfolio selection problems. In: Bamberg, G., Spreemann, K. (Eds.), Risk and Capital. Springer-Verlag, Berlin/New York, pp. 74–87.
Kim, T.S., Omberg, E., 1996. Dynamic nonmyopic portfolio behavior. Review of Financial Studies 9, 141–161.
Kingsland, L., 1982. Projecting the financial condition of a pension plan using simulation analysis. Journal of Finance 37, 577–584.
Klaassen, P., 1997. Discretized reality and spurious profits in stochastic programming models for asset/liability management. European Journal of Operational Research 101, 374–392.

Klaassen, P., 1998. Financial asset-pricing theory and stochastic programming models for asset-liability management: A synthesis. Management Science 44, 31–48.
Konno, H., Kobayashi, K., 1997. An integrated stock-bond portfolio optimization model. Journal of Economic Dynamics and Control 21, 1427–1444.
Konno, H., Yamazaki, H., 1991. Mean-absolute deviation portfolio optimization model and its applications to Tokyo stock market. Management Science 37 (5), 519–531.
Koskosides, Y., Duarte, A., 1997. A scenario-based approach for active asset allocation. The Journal of Portfolio Management, 74–85, Winter.
Kouwenberg, R., 2001. Scenario generation and stochastic programming models for asset liability management. European Journal of Operational Research 134, 51–64.
Kusy, M.I., Ziemba, W.T., 1986. A bank asset and liability management model. Operations Research 34, 356–376.
Liu, J., 1999. Portfolio selection in stochastic environments. Working paper, Stanford University, Stanford, US.
Maranas, C.D., Androulakis, I.P., Floudas, C.A., Berger, A.J., Mulvey, J.M., 1997. Solving long-term financial planning problems via global optimization. Journal Economic Dynamics and Control 21, 1405–1425.
Markowitz, H.M., 1952. Portfolio selection. Journal of Finance 7, 77–91.
Merton, R.C., 1969. Lifetime portfolio selection under uncertainty: The continuous-time case. Review of Economics and Statistics 51, 247–257.
Merton, R.C., 1971. Optimum consumption and portfolio rules in a continuous-time model. Journal of Economic Theory 3, 373–413.
Merton, R.C., 1990. Continuous-Time Finance. Blackwell, Cambridge, MA.
Mossin, J., 1968. Optimal multiperiod portfolio policies. Journal of Business 41, 215–229.
Mulvey, J.M., 1993. Incorporating transaction costs in models for asset allocation. In: Zenios, S.A. (Ed.), Financial Optimization. Cambridge University Press, Cambridge, UK, pp. 243–259.
Mulvey, J.M., 1996. Generating scenarios for the towers Perrin investment system. Interfaces 26, 1–15.
Mulvey, J.M., Vladimirou, H., 1992. Stochastic network programming for financial planning problems. Management Science 38, 1643–1664.
Mulvey, J.M., Zenios, S.A., 1994. Capturing the correlations of fixed-income instruments. Management Science 40, 1329–1342.
Mulvey, J.M., Gould, G., Morgan, C., 2000. An asset and liability management system for Towers Perrin–Tillinghast. Interfaces 30 (1), 96–114.
Naik, V., 1995. Finite state securities market models and arbitrage. In: Jarrow, R., et al. (Eds.), Handbooks in OR&MS. Elsevier Science, Amsterdam, pp. 31–64.
Nielsen, S.S., Zenios, S.A., 1996a. Solving multistage stochastic network programs on massively parallel computers. Mathematical Programming 75, 227–250.
Nielsen, S.S., Zenios, S.A., 1996b. A stochastic programming model for funding single premium deferred annuities. Mathematical Programming 75, 177–200.
Pflug, G.Ch., 2001. Optimal scenario tree generation for multiperiod financial planning. Mathematical Programming 89, 251–271.
Reddington, F.M., 1952. Review of the principles of life-office valuations. Journal of the Institute of Actuaries 78, 286–340.
Riskmetrics, 1996. Technical document, fourth ed., J.P. Morgan, New York, NY.
Rockafellar, R.T., Uryasev, S., 2000. Optimization of conditional Value-at-Risk. The Journal of Risk 2 (3), 21–41.
Rudolf, M., Ziemba, W.T., 2004. Intertemporal surplus management. Journal of Economic Dynamics and Control 28, 975–990.
Samuelson, P.A., 1969. Lifetime portfolio selection by dynamic stochastic programming. Review of Economics and Statistics 51, 239–246.
Sharpe, W.F., Tint, L.G., 1990. Liabilities—a new approach. Journal of Portfolio Management, 5–10, Winter.
Shtilman, M.S., Zenios, S.A., 1993. Constructing optimal samples from a binomial lattice. Journal of Information and Optimization Sciences 14, 1–23.

Sortino, F.A., van der Meer, R., 1991. Downside risk—capturing whats at stake in investment situations. Journal of Portfolio Management 17, 27–31.

Topaloglou, N., Vladimirou, H., Zenios, S.A., 2002. CVaR models with selective hedging for interactional asset allocation. Journal of Banking and Finance 26 (7), 1535–1561.

Wets, R.J.-B., 1989. Stochastic programming. In: Nemhauser, G.L., Rinnooy Kan, A.H.G., Todd, M.J. (Eds.), Handbooks in Operations Research and Management Science, vol. 1. North-Holland, Amsterdam, pp. 573–629.

Winklevoss, H.E., 1982. Plasm: Pension liability and asset simulation model. Journal of Finance 37, 585–594.

Worzel, K.J., Vassiadou-Zeniou, C., Zenios, S.A., 1994. Integrated simulation and optimization models for tracking fixed-income indices. Operations Research 42 (2), 223–233, March–April.

Xia, Y., 2000. Learning about predictability: The effects of parameter uncertainty on dynamic asset allocation. Journal of Finance 56, 585–594.

Zenios, S.A., 1991. Massively parallel computations for financial modeling under uncertainty. In: Mesirov, J. (Ed.), Very Large Scale Computing in the 21st Century. SIAM, Philadelphia, PA, pp. 273–294.

Zenios, S.A. (Ed.), 1993. Financial Optimization. Cambridge University Press, Cambridge, UK.

Zenios, S.A., 1995. Asset/liability management under uncertainty for fixed-income securities. Annals of Operations Research 59, 77–98; Reprinted in: Ziemba, W.T., Mulvey, J.M. (Eds.), World Wide Asset and Liability Modeling. Cambridge University Press, Cambridge, UK, 1998.

Zenios, S.A., 2006. Practical Financial Optimization: Decision Making for Financial Engineers. Basil Blackwell, Oxford, in press.

Zenios, S.A., Holmer, M., McKendall, R., Vassiadou-Zeniou, C., 1998. Dynamic models for fixed-income portfolio management under uncertainty. Journal of Economic Dynamics and Control 22, 1517–1541.

Ziemba, W.T., Mulvey, J.M. (Eds.), 1998. Worldwide Asset and Liability Modeling. Cambridge University Press, Cambridge, UK.

Ziemba, W.T., Vickson, R.G. (Eds.), 1975. Stochastic Optimization Models in Finance. Academic Press, San Diego, CA.

This page intentionally left blank

Chapter 7

BOND PORTFOLIO MANAGEMENT VIA STOCHASTIC PROGRAMMING

M. BERTOCCHI and V. MORIGGIA

Department of Mathematics, Statistics, Computer Science and Applications, Bergamo University, Via dei Caniana 2, 24127 Bergamo, Italy
e-mails: marida.bertocchi@unibg.it; vittorio.moriggia@unibg.it

J. DUPAČOVÁ

Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics of Charles University, Sokolovská 83, CZ-186 75 Prague 8, Czech Republic
e-mail: dupacova@karlin.mff.cuni.cz

Contents

Abstract 306
1. Introduction 307
2. The bond portfolio management model 311
3. Input data 315
4. Scenario reduction and scenario tree construction 320
5. Numerical results 321
6. Stress testing via contamination: Add worst-case scenarios 325
7. Conclusions 334
Acknowledgements 335
References 335

Handbook of Asset and Liability Management, Volume 1
Edited by S.A. Zenios and W.T. Ziemba
Copyright © 2006 Elsevier B.V. All rights reserved
DOI: 10.1016/S1872-0978(06)01007-6

Abstract

Stochastic programming is a tool to support bond portfolio management decisions. For a successful application of the stochastic programming methodology, one must choose an adequate model, asses its parameters, generate sensible input scenarios or scenario tree, solve the scenario-based problem using an optimization software and validate the results. Formulation of two-stage and three-stage models for bond portfolio optimization are discussed and computational results displayed. Comparisons of monthly and quarterly time discretizations and for various topologies of the input scenario trees are presented. Black–Derman–Toy binomial lattice calibrated from market data is used to generate scenarios and scenario reduction and scenario construction methods are applied using GAMS. The contamination technique is exploited to quantify the impact of including additional (stress or out-of-sample) scenarios and/or additional stages to an already selected scenario tree.

JEL classification: C61, E47, G11

1. Introduction

The problem is to preserve the value of a bond portfolio of an institutional investor over time. This is a problem of allocation and management of resources, not of high frequency trading. It may include additional features, e.g., presence of fixed or uncertain external inflows or outflows in the future or a required balance between assets and liabilities. There are various options concerning the choice of an appropriate model. Historically, the asset-liability portfolio models have focused on duration matching, immunization techniques, cf. chapter by de La Grandville (2006) in this volume, and on cash flow matching, dedication models which try to match the assets and liability flows at every time period, cf. Hiller and Eckstein (1994); Zenios (1993). Whereas duration matching strategies require an active portfolio management, dedication models capture the time evolution of cash flows and admit a more passive management style. Moreover, while the dedication models have to face the reinvestment risk problem, the duration matching models addressed it by an implicitly included continuously rebalancing strategy. To reflect the uncertainties in prices, interest rates, etc., simple stochastic analogs of these two approaches were developed, cf. Hiller and Eckstein (1994); Hiller and Schaack (1990); Shapiro (1988), and the tracking models by Dembo (1993). They make use of alternative scenarios of evolution of the uncertain input parameters.

Why not to rely on the duration based immunization models? A good answer is the following quotation, cf. Kahn (1991):

> "Many years ago, bonds were boring. Returns were small and steady. Fixed income risk monitoring consisted in watching duration and avoiding low qualities. But as interest-rate volatility has increased and the variety of fixed income instruments has grown, both opportunities and dangers have flourished..."

Yield curves are not flat, do not move in a parallel way, the interest rates are not constant, investment in long maturity bonds requires an active rebalancing strategy and the notion of duration for more complex financial instruments is not obvious. The scenario-based static models mentioned above were a significant step forward for bond portfolio management. The next step are applications of multiperiod and multistage stochastic programs which capture both the uncertainties and dynamic features in a more adequate way.

To keep the presentation as simple as possible, we consider only one random factor to be included—the evolution of the short interest rate over time.

Given a sequence of equilibrium future short term interest rates r_t valid for the time interval $(t, t+1]$, $t = 0, \ldots, \tau - 1$, the fair price of the jth bond at time t just after the coupon was paid equals the present value of the stream of cash flows f_{jl}, $l = t+1, \ldots, \tau$, generated by this bond in subsequent time instances discounted to t

$$P_{jt}(\boldsymbol{r}) = \sum_{l=t+1}^{\tau} f_{jl} \prod_{h=t}^{l-1} (1 + r_h)^{-1} \tag{1}$$

where τ is greater or equal to the time to maturity.

The time points need not coincide with the dates of coupon payments. The sequence of the future short term rates r_l that determines the prices (1) is not known precisely, but prescribed ad hoc or modeled in a probabilistic way. The cash flows f_{jl} need not be known with certainty; this is for instance the case of indexed bonds, bonds with options or default. Hence, the formula (1) should be extended for the accrued interest and revised to take into account the effect of options and other risks related with the jth bond. The resulting selling and purchasing prices reflect transaction costs and the bid/ask spread.

The future interest rates are not known with certainty, they are treated as random. We assume that their probability distribution is well approximated by a discrete probability distribution with a finite number of τ-dimensional vectors $\boldsymbol{r}^s$ consisting of the short rates r_t^s, $t = 0, \ldots, \tau - 1$, $s = 1, \ldots, S$, with probabilities $p_s > 0$, $s = 1, \ldots, S$, $\sum_s p_s = 1$. The rate r_0 valid in the first period is known. Vectors $\boldsymbol{r}^s$ are atoms of the discrete probability distribution P of interest rates and are called *scenarios*.

Scenario-based stochastic programs are formulated to reflect the characteristic features of the bonds included into the portfolio, the time discretization, constraints and the goal to be achieved at a given planning horizon. Their application requires an adequate representation of the dynamic aspects and of the involved uncertainties, i.e., of the probability distribution of the random factors, in our case, of the interest rates. Considering the solved problem, extension for additional sources of randomness may be important. This is the case of credit spreads for corporate bonds, cf. Nielsen and Ronn (1997); Bertocchi, Giacometti and Zenios (2005), embedded options in callable bonds, cf. Vassiadou-Zeniou and Zenios (1996), prepayments of mortgages, cf. Kang and Zenios (1992) for portfolio of mortgage-backed securities, exchange rates for portfolio of international bonds, cf. Jamshidian and Zhu (1997), inflation, cf. Mulvey and Thorlacius (1998) and further macroeconomic or demographic factors.

In applications of stochastic programming for portfolio optimization one can exploit standardized types of stochastic programming models (e.g., two-stage and multistage stochastic programs with recourse, models with individual and joint probabilistic constraints, integer stochastic programs) and software systems, see, e.g., Birge and Louveaux (1997); Dupačová, Hurt and Štěpán (2002, Part II); Kall and Wallace (1994); Prékopa (1995); Kouwenberg and Zenios (2006); Wallace and Ziemba (2005). We refer to chapter by Kouwenberg and Zenios (2006) in this volume for comparisons of stochastic programming with other optimization methods applied in finance and for a survey of successful applications.

The constraints follow the cash flow accounting rules and appear in the form of (time and scenario dependent) mostly *linear* constraints on cash and inventory balance and regulatory constraints. They are complemented by *nonanticipativity constraints* which capture an additional important feature of the decision process: *one is allowed to use only the available* (*past*) *information and updates the decision when an additional information gets revealed*.

Accordingly, for *scenario-based multiperiod and multistage stochastic programs* the input is usually in the form of a fan of scenarios or of a scenario tree, respectively. The

nonanticipativity constraints on decisions may enter implicitly or in an explicit way. In both cases decisions based on the same history (i.e., on an identical part of several scenarios) are forced to be equal. With the explicit inclusion of the nonanticipativity constraints, the scenario-based multiperiod and multistage stochastic program with linear constraints can be written as the large-scale deterministic program

$$\max_{\mathcal{X}_0 \cap \mathcal{C}} \left\{ \sum_s p_s u^s(\boldsymbol{x}^s) \mid \boldsymbol{A}^s \boldsymbol{x}^s = \boldsymbol{b}^s,\ s = 1, \ldots, S \right\}. \tag{2}$$

Here $\mathcal{X}_0$ is a set of "hard" constraints such as nonnegativity conditions, $\mathcal{C}$ is defined by the nonanticipativity constraints, u^s is the performance measure when scenario s occurs (with probability p_s) and $\boldsymbol{x}^s$ is the corresponding decision vector. In the case of two-stage stochastic programs the nonanticipativity constraints correspond to the requirement that the first-stage decisions must be scenario independent, i.e., that they are equal for all scenarios.

The implicit inclusion of nonanticipativity constraints leads to the *arborescent or nodal* formulation of the stochastic program: Each node of the scenario tree corresponds to a history of the random process up to a certain time t, a stage at which decisions may be taken. The last decision point (stage) is related to the chosen planning horizon. Assuming discrete-time data processes, the nodes may be numbered as $n = 0, \ldots, N$ with index $n = 0$ assigned to the root—the only node at stage $t = 1$. For $t = 1, \ldots, T - 1$ nodes at stage $t + 1$ are elements of the set $\mathcal{N}_t$. The (unique) predecessor of node $n \in \mathcal{N}_t$ at the stage t is marked as $\hat{n}$. Let $\mathcal{D}(n)$ be the set of descendants of the node $n \in \mathcal{N}_t$; the elements of $\mathcal{D}(n)$ are then the nodes from $\mathcal{N}_{t+1}$ which can be reached from the node n. Sets $\mathcal{D}(n)$ may contain only one element. The extremal case when there are multiple descendants only of the root ($n = 0$) relates to the *fan of scenarios* and corresponds to two-stage (multiperiod) problems. The probability of reaching the node $n \in \mathcal{N}_t$ is p_{tn}. For planning horizon T, nodes belonging to the set $\mathcal{N}_{T-1}$ are called *leaves* and a scenario corresponds to a path from the root to some $n \in \mathcal{N}_{T-1}$. Given scenario probabilities p_{Tn} a path probability can be assigned to each node by a recursion.

At each node of the scenario tree a decision $\boldsymbol{x}_n$ is taken. Constraints of (2) are rewritten as

$$\begin{aligned} &\boldsymbol{W}_0 \boldsymbol{x}_0 = \boldsymbol{b}_0, \quad \boldsymbol{x}_0 \in \mathcal{X}_0, \\ &\boldsymbol{T}_n \boldsymbol{x}_{\hat{n}} + \boldsymbol{W}_n \boldsymbol{x}_n = \boldsymbol{b}_n, \quad \boldsymbol{x}_n \in \mathcal{X}_n,\ n \in \mathcal{N}_t,\ t = 1, \ldots, T - 1. \end{aligned} \tag{3}$$

The matrix $\boldsymbol{W}_0$ and the right-hand side vector $\boldsymbol{b}_0$ in the first-stage constraints are fixed in advance, matrices $\boldsymbol{W}_n$, $\boldsymbol{T}_n$ and vectors $\boldsymbol{b}_n$ result from the history preceding the node n. For all nodes, sets $\mathcal{X}_n$ are defined by separate constraints on $\boldsymbol{x}_n$. In this nodal formulation the objective function of the *split variable* form (2) becomes

$$\sum_{n \in \mathcal{N}_{T-1}} p_{Tn} u^n(\boldsymbol{x}_n). \tag{4}$$

Besides the formulation of goals and constraints and identification of the driving random process, building a scenario-based multiperiod or multistage stochastic program requires specification of the horizon, stages and generation of the input in the form of a scenario fan or a scenario tree. In a majority of cases, the horizon and the stages are given. The horizon may be often determined by the nature of the decision problem. It may be tied to a fixed date, e.g., to the end of the fiscal year, to a date related with the annual Board of Directors' meeting, or to the deadline for repayment of the debt. Another possibility is to consider the horizon connected with a time interval of a fixed (possibly even infinite) length, given for instance by the periodicity of the underlying random process.

The problem may be solved repeatedly, with new horizons, taking into account just the already achieved state of the system. To guarantee the possibility of such continuation, the models are usually extended for additional constraints and/or terms in the objective function to reduce the end effects.

Rolling forward after the T-stage problem has been solved, a first-stage decision accepted and a new information obtained means to solve a subsequent $(T-1)$-stage stochastic program with a reduced number of stages or another T-stage problem with the initial state of the system determined by the applied first-stage decision and an observed realization of the random factor and using the random process shifted in time.

Also selection of stages results sometimes from the problem formulation (e.g., the dates of maturity of bonds, cf. Frauendorfer and Marohn (1998), or expiration dates of options) but more frequently, stages are fixed ad hoc, by application of heuristic rules and/or experience and regarding software and computer facilities.

Hence, for an already chosen horizon, the crucial step is to relate the time instants and stages. To use multiperiod two-stage model or to assign one stage to each of discretization points are two extreme cases. One must design then sensible and tractable scenario generation procedures taking into account the chosen type of model, the level of the existing information, software and hardware possibilities. Finally, a validation of results is necessary: an approximate stochastic program is mostly solved instead of the underlying "true" decision problem and compromises between the size of the resulting problem and the desired precision of the results are accepted. A detailed analysis of the origin and of the initial structure of the solved problem may be exploited to aggregate the stages, to prune the tree or to extend it for other out-of-sample scenarios or branches. It is even possible to test the influence of including additional stages, e.g., using the contamination approach, and to use it in stress testing.

Various stochastic models of short interest rates may be used to generate the interest rate scenarios. We follow the Black, Derman and Toy (1990) model (BDT); in the sector of fixed income securities the BDT model is close to the industry standard. Its advantage is, inter alia, that its calibration, based on market data, is separated from sampling techniques suggested to create a scenario tree. Hence, there are at disposal many scenarios—all paths through the BDT binomial lattice, see Section 3. Their number depends on the chosen time discretization and on the maturities of the bonds to be included in the portfolio. These interest rate scenarios are used to price the bonds ac-

cording (1) and they extend behind the considered horizon of the stochastic program. In two-stage stochastic programs the input coefficients are based directly on these scenarios whereas for application of multistage models, scenario trees must be generated. See, e.g., Dupačová, Consigli and Wallace (2000) and references therein for a survey of clustering, sampling and fitting techniques.

Several papers discuss the model performance in dependence on transaction costs, spread, etc., sensitivity of results on the structure of the model, on selection of representative scenarios of interest rates and out-of-sample behavior, influence of errors due to input data and robustness of the first-stage decision with respect to perturbations of the estimated term structure, see, e.g., Bertocchi, Dupačová and Moriggia (2000, 2006); Dupačová (2000); Dupačová and Bertocchi (2001); Dupačová, Bertocchi and Moriggia (1998).

We focus on comparisons of results for monthly and quarterly time discretizations and for various topologies of the input scenario trees. The scenario reduction and scenario construction methods, cf. Dupačová, Gröwe-Kuska and Römisch (2003); Gröwe-Kuska, Heitsch and Römisch (2003), are tested for the two-stage and multistage version of the bond portfolio management model introduced in Section 2. Similarly as in Dupačová, Bertocchi and Moriggia (1998), the contamination technique is applied to quantify the impact of including additional (stress or out-of-sample) scenarios and/or additional stages to the already selected scenario tree.

2. The bond portfolio management model

Bond portfolio management was one of the first applications of stochastic programming in finance, cf. Bradley and Crane (1972). Several models have been proposed and applied, e.g., Dupačová, Bertocchi and Moriggia (1998); Frauendorfer and Marohn (1998); Golub et al. (1995). We consider a bond portfolio management problem of an institutional investor to maximize the expected performance of the portfolio over time. As in Golub et al. (1995), rebalancing portfolio at certain time instants is allowed. This is a generalization of the stochastic dedicated bond portfolio problem introduced by Shapiro (1988), and also Hiller and Eckstein (1994).

Let

$j = 1, \ldots, J$ indices of the bonds and T_j the dates of their maturities; $\tau = \max_j T_j$;
$t = 0, \ldots, T$ the discretization of the planning horizon;
$b_j \geqslant 0$ the initial holdings of bond j;
b_0 the initial holding in riskless asset;
f_{jt}^s cash flow generated under scenario s from the unit quantity of bond j at time t;
ξ_{jt}^s and ζ_{jt}^s are the selling and purchasing prices of bond j at time t for scenario s obtained from the corresponding fair prices (1) adding the accrued interest A_{jt}^s and subtracting or adding scenario independent transaction costs and spread; the initial prices ξ_{j0} and ζ_{j0} are known constants, i.e., scenario independent;

L_t is an external cash flow at time t;
x_j/y_j are face values of bond j purchased/sold at the beginning of the planning period, i.e., at $t = 0$;
z_{j0} is the quantity of bond j held in portfolio after the initial decisions x_j, y_j have been made;
$x^s_{jt}, y^s_{jt}, z^s_{jt}$ are the corresponding decisions at time t under scenario s.

The first-stage decision variables x_j, y_j, z_{j0} are nonnegative

$$y_j + z_{j0} = b_j + x_j \quad \forall j, \tag{5}$$

$$y_0^+ + \sum_j \zeta_{j0} x_j = b_0 + \sum_j \xi_{j0} y_j \tag{6}$$

where the nonnegative variable y_0^+ denotes the surplus. In the first stage, no borrowing is permitted. This limitation together with the assumed *positive market value of the initial portfolio*, $b_0 + \sum_j \xi_{j0} b_j > 0$ implies that *the set $\mathcal{X}_0$ of the feasible first-stage solutions* $(y_0^+, x_j, y_j, z_{0j}\ \forall j)$ *is nonempty and bounded*. Notice, that it does not depend on scenarios. The same property holds true also in case that a *restricted borrowing* possibility in the first stage is permitted.

The second-stage decisions on rebalancing the portfolio, borrowing or reinvestment of the surplus depend on individual scenarios. They have to fulfil constraints on conservation of holdings in each bond at each time period and for each of scenarios

$$z^s_{jt} + y^s_{jt} = z^s_{j,t-1} + x^s_{jt} \quad \forall j, s, t \geqslant 1, \tag{7}$$

$$\sum_j \xi^s_{jt} y^s_{jt} + \sum_j f^s_{jt} z^s_{j,t-1} + (1 - \delta_1 + r^s_{t-1}) y^{+s}_{t-1} + y^{-s}_t = L_t + \sum_j \zeta^s_{jt} x^s_{jt} + (1 + \delta_2 + r^s_{t-1}) y^{-s}_{t-1} + y^{+s}_t, \ \forall s, t, \tag{8}$$

with nonnegativity of all variables and with $y_0^{-s} = 0, y_0^{+s} = y_0^+, z^s_{j0} = z_{j0}\, \forall s, j$. Variables y^{+s}_t, y^{-s}_t describe the (unlimited) lending/borrowing possibilities for period t under scenario s and the nonnegative spreads δ_1, δ_2 are model parameters to be fixed. Nonzero values of δ_1 account for the difference between the returns for bonds and for cash. A natural assumption is $\delta_2 > 0$, i.e., there is a positive cost of borrowing.

The optimization problem is maximizing the expected utility of the final wealth at the planning horizon T

$$\sum_s p_s U(W^s_T) \tag{9}$$

subject to (5)–(8), nonnegativity constraints on all variables and with

$$W^s_T = \sum_j \xi^s_{jT} z^s_{jT} + y^{+s}_T - \alpha y^{-s}_T. \tag{10}$$

The multiplier $\alpha \geqslant 1$ is fixed according to the problem area. For example, large values of α play a role of penalty for borrowing at the end of the accounting or planning period.

Because of the possibility of reinvestment and of unlimited borrowing, the second-stage constraints (7), (8) have always a nonnegative solution. It is then a *multiperiod two-stage scenario-based stochastic program with relatively complete recourse* with linear constraints and nonlinearities in the objective function. The existence of optimal solution is guaranteed for a large class of utility functions that are *increasing and concave*. Moreover, for strict inequalities between the purchasing and selling prices, $\xi_{j0} < \zeta_{j0}$, $\xi^s_{jt} < \zeta^s_{jt}$ $\forall j, t, s$, and $\delta_1 \geqslant 0$, $\delta_2 > 0$, the optimal solutions satisfy

$$y_j x_j = 0 \,\forall j, \quad y^s_{jt} x^s_{jt} = 0 \,\forall j, t, s, \quad y^{+s}_t y^{-s}_t = 0 \,\forall t, s. \tag{11}$$

At optimality there is no unnecessary rebalancing and borrowing, which is a natural property.

Provided that an initial rebalancing strategy determined by feasible, scenario independent first-stage decisions x_j, y_j, $\forall j$ has been accepted, the second-stage scenario-dependent decisions must be made in an optimal way regarding the goal of the model—to maximize the final wealth subject to constraints on conservation of holdings, with possible rebalancing the portfolio

$$\text{maximize} \quad W^s_T := \sum_j \xi^s_{jT} z^s_{jT} + y^{+s}_T - \alpha y^{-s}_T \tag{12}$$

subject to (7), (8) and nonnegativity constraints.

The main outcome is the optimal value of the objective function (9) and the optimal values of the first-stage variables x_j, y_j (and y^+_0, z_{j0}) for all j. The optimal values of the second-stage variables contribute to evaluation of the objective function (9). In a dynamic setting, the first-stage decision is applied at the beginning of the first period, and at the beginning of the next period, the model is solved again for the changed input information on holdings and on scenarios of interest rates; this is the *rolling forward* approach in the context of two-stage stochastic programming models, see Kusy and Ziemba (1986) for an explanation of this idea.

The model may be extended for further constraints, such as allocation group restrictions (box constraints on holdings in specified assets categories—cash, medium term bonds, long bonds, etc.) or benchmark targets concerning wealth and/or rate of return, cf. Cariño and Turner (1998). Another suggestion, see Messina and Mitra (1996), is to limit the gain or loss from intertemporal rebalancing by constraints

$$\sum_j |\xi^s_{j,t-1} - \xi^s_{jt}| \, y^s_{jt} \leqslant V_t \quad \forall s, t, \tag{13}$$

or to limit the portfolio's duration (Mulvey and Zenios, 1994). As an example, we require that

$$D^* - \Delta \cdot D^* \leqslant \sum_{j=1}^{J} z_{0j} D_j \leqslant D^* + \Delta \cdot D^* \tag{14}$$

where

$$D_j = \sum_{t=1}^{\tau} \frac{t f_{jt}}{(1+\hat{r}_j)^{t+1}}, \quad j = 1, \dots J, \tag{15}$$

is the dollar duration of bond j, $\hat{r}_j$ its yield to maturity,

$$D^* = \sum_{j=1}^{J} b_{0j} D_j, \tag{16}$$

is the initial portfolio duration and Δ is the chosen tolerance parameter.

Another possibility is to design a *multistage model*. This means that another *true* rebalancing based on a newly revealed information is allowed in one or more later periods. Such information can be the recognition of increasing or decreasing interest rates, the outcome concerning an exercise of option, etc. Then it is necessary to design also the input in the form of a scenario tree along with the path and arc probabilities and to incorporate the nonanticipativity constraints.

Following Dupačová (2000); Bertocchi, Dupačová and Moriggia (2006) we now introduce a multistage reformulation.

Assume that r_0 is the known interest rate for the first period (0, 1], r_1^s, $1 \leqslant s \leqslant S$, the considered realizations of the interest rate valid for the period (1, 2], with probabilities $p_s > 0$, $\sum_s p_s = 1$, and denote $\mathcal{D}(s)$ the set of descendants of scenario s—the set of rates for the subsequent time periods $r_t^{s\sigma}$ with conditional (arc) probabilities $\pi_\sigma > 0$, $\sigma \in \mathcal{D}(s)$, $\sum_{\sigma \in \mathcal{D}(s)} \pi_\sigma = 1$. The expected optimal outcome of the second rebalancing strategy $\boldsymbol{x}_1^s$, $\boldsymbol{y}_1^s(z_1^s, y_1^{+s})$ (in its dependence on a feasible first-stage decision $\boldsymbol{x}$, $\boldsymbol{y}$) is evaluated along the branch of scenarios $r_t^{s\sigma}$, $2 \leqslant t \leqslant T-1$, $\sigma \in \mathcal{D}(s)$, emanating from r_1^s and equals the *optimal value* $W_T^s(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}_0, y_0^+)$ of the following two-stage stochastic linear program (compare with (5)–(10)):

$$W_T^s(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}_0, y_0^+) := \max \sum_{\sigma \in \mathcal{D}(s)} \pi_\sigma \left[\sum_j \xi_{jT}^{s\sigma} z_{jT}^{s\sigma} + y_T^{+s\sigma} - \alpha y_T^{-s\sigma} \right] \tag{17}$$

subject to

$$y_{j1}^s + z_{j1}^s = z_{j0} + x_{j1}^s \quad \forall j, \tag{18}$$

$$L_1 + y_1^{+s} + \sum_j \zeta_{j1}^s x_{j1}^s = (1 - \delta_1 + r_0) y_0^+ + \sum_j \xi_{j1}^s y_{j1}^s + \sum_j f_{j1} z_{j0} + y_1^{-s}, \tag{19}$$

$$z_{jt}^{s\sigma} + y_{jt}^{s\sigma} = z_{j,t-1}^{s\sigma} + x_{jt}^{s\sigma} \quad \forall j,\ 2 \leqslant t \leqslant T,\ \sigma \in \mathcal{D}(s), \tag{20}$$

$$\sum_j \xi_{jt}^{s\sigma} y_{jt}^{s\sigma} + \sum_j f_{jt} z_{j,t-1}^{s\sigma} + (1 - \delta_1 + r_{t-1}^{s\sigma}) y_{t-1}^{+s\sigma} + y_{t-1}^{+s\sigma}$$
$$= L_t + \sum_j \zeta_{jt}^{s\sigma} x_{jt}^{s\sigma} + y_t^{+s} + (1 + \delta_2 + r_{t-1}^{s\sigma}) y_{t-1}^{-s\sigma}, \quad 2 \leqslant t \leqslant T,\ \sigma \in \mathcal{D}(s), \tag{21}$$

with $r_1^{s\sigma} = r_1^s$, $y_1^{+s\sigma} = y_1^{+s}$, $y_1^{-s\sigma} = y_1^{-s}$, $z_{j1}^{s\sigma} = z_{j1}^s$ $\forall j, \sigma \in \mathcal{D}(s)$ and under nonnegativity of all variables.

The stochastic program is

$$\text{maximize} \quad \sum_s p_s U\big(W_T^s(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}_0, y_0^+)\big) \tag{22}$$

subject to (5)–(6) and nonnegativity constraints.

The prices $\xi_{jt}^{s\sigma}$, $\zeta_{jt}^{s\sigma}$ are obtained from the fair prices $P_{jt}(\boldsymbol{r}^{s\sigma})$ computed according to (1) for $t \geqslant 2$ and $\sigma \in \mathcal{D}(s)$. The fair prices P_{j1} follow by the expectation hypothesis: they are equal to the discounted expected value of the sum of cash flows f_{j2} due at $t = 2$ and fair prices $P_{j2}(\boldsymbol{r}^{s\sigma})$:

$$P_{j1}^s = (1 + r_1^s)^{-1} \sum_{\sigma \in \mathcal{D}(s)} \pi_\sigma \big(f_{j2} + P_{j2}(\boldsymbol{r}^{s\sigma})\big). \tag{23}$$

This is the way how to get the required tree structure of the model coefficients. In comparison with the initial two-stage problem (5)–(6), (9) with W_T^s defined by (10), using the above arborescent form of the three-stage problem results in a decrease of the size of the problem as to the number of variables.

Based on the tree structure of coefficients, the three-stage problem (5)–(6), (22) with W_T^s defined by (17)–(21) can be also written in the split variable form which formally corresponds to the two-stage formulation complemented for explicit nonanticipativity constraints. This formulation evidently means an extension of the size of the deterministic problem to be solved.

Similar formulas for prices apply also for other instances of scenario tree structures. However if the branchings appear after several time periods one has to apply specific rules for the interstage portfolio management. This is mostly the buy-and-hold strategy with accumulation of cash and coupons according to interest rates. When no bond matures within the time horizon, only coupons are transferred to cash account. Another possibility would be to allow for reinvesting coupons into the security which provided them.

In the numerical experiments we focus on the problem of a risk neutral institutional investor (i.e., with a linear utility function) whose task is to preserve the value of the bond portfolio. Hence, no external cash flows are considered, so that $L_t = 0$, $y_t^{-s} = 0$ $\forall t, s$. We assume that the portfolio consists of default free, liquid bonds with maturities $T_j > T$, all cash flows are after tax, the transaction costs and bid/ask spread are fixed constants.

3. Input data

The initial portfolio for October 3, 1994 was designed by experts, see Table 1. We set the planning horizon to 1 year and use monthly or quarterly time discretization. The considered date allows us to include also puttable bonds in the portfolio; moreover,

Table 1
Portfolio composition on October 3rd 1994

Bonds	Qt	Coupon	Payment dates	Exercise	Redemp.	Maturity
BTP36658	10	3.9375	01Apr & 01Oct		100.187	01Oct96
BTP36631	20	5.0312	01Mar & 01Sep		99.531	01Mar98
BTP12687	15	5.2500	01Jan & 01Jul		99.231	01Jan02
BTP36693	10	3.7187	01Aug & 01Feb		99.387	01Aug04
BTP36665	5	3.9375	01May & 01Nov		99.218	01Nov23
CTO13212	20	5.2500	20Jan & 20Jul	20Jan95	100.000	20Jan98
CTO36608	20	5.2500	19May & 19Nov	19May95	99.950	19May98

none of the bonds expires within the investment horizon. The initial portfolio market value is equal to 10 484.55 and the initial portfolio's dollar duration is 36 407.7.

The results use the Black–Derman–Toy (BDT) model of Black, Derman and Toy (1990) for generation of interest rate scenarios. The interest rates r_t^s valid in the interval $(t, t+1]$, $t > 0$, are

$$r_t^s = r_{t0} k_t^{i_t(s)} \tag{24}$$

where $i_t(s) = \sum_{l=1}^{t} w_l^s$ is the scenario dependent position on the lattice that equals the number of up moves, coded by $w_l^s = 1$, which occur at time points $l = 1, \ldots, t$ for the given scenario s.

The scenario independent quantities r_{t0}, the lowest short rate that may occur for period $(t, t+1]$, and the lattice volatilities k_t are obtained by calibration of the BDT model by the (estimated) market structure of a given day, mostly of October 3, 1994. These steps lead to the fitted binomial lattice which provides different 2^τ scenarios of interest rates where τ is the longest bond maturity in the portfolio.

Hence, to price the longest bond BTP36665 of the portfolio from Table 1 would mean to fit the lattice over the 29 years period. We used a numerically less demanding procedure—which is better supported by the observed market data—to fit the lattice using the term structure up to 5 years maturity and to price bonds with longer maturities using constant interest rates in time periods after the 5 years cutting point; see Figure 1 for the scheme and Figure 2 for the resulting lattice of interest rates valid in the four quarters of the yearly planning period.

With quarterly and monthly discretizations of the one year planning horizon the full lattices up to the horizon will provide $2^4 = 16$ and $2^{12} = 4096$ equiprobable scenarios, respectively. Figure 2 relates to a part of the BDT lattice corresponding to $2^4 = 16$ scenarios for the 4 quarters of the quarterly binomial lattice, numbered as [0], ..., [15]. The scenarios are obtained by splitting the nodes of this lattice and are mapped to the sequence of the corresponding numbers of nodes, one node per scenario and quarter; see Figure 3 for the fan of scenarios and nodes, a starting point for the two-stage multiperiod bond portfolio management problem. The extremal scenarios [15] "up, up, up, up" and

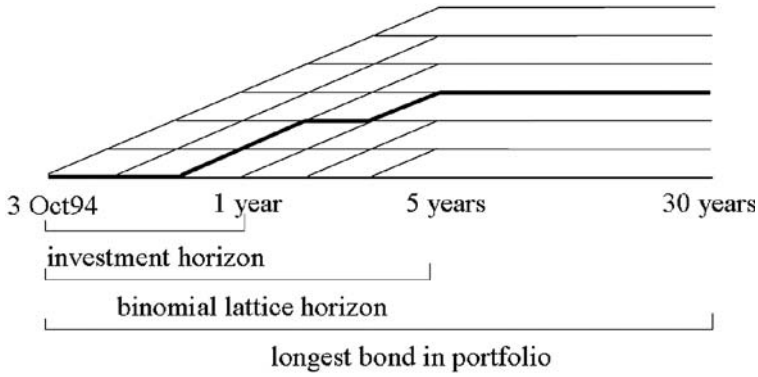


Fig. 1. Binomial lattice.

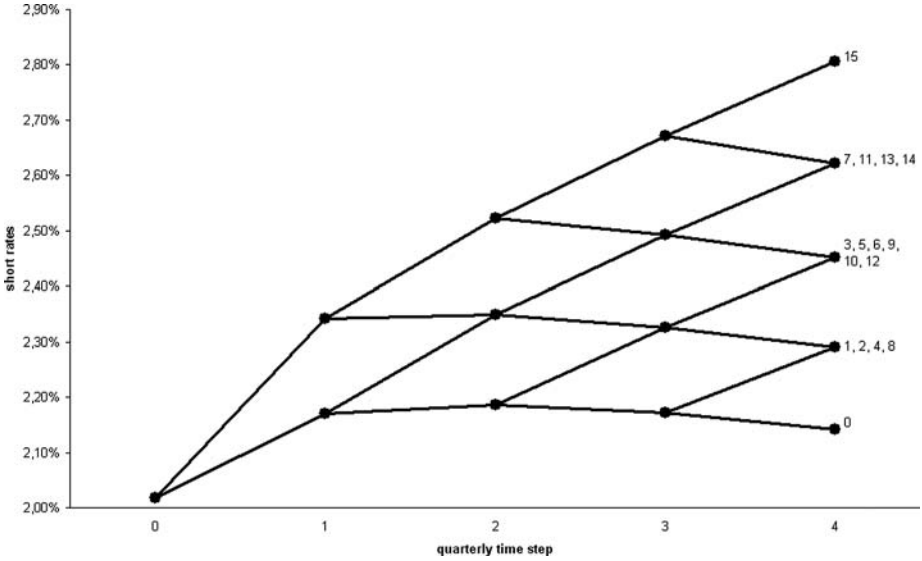


Fig. 2. Scenarios 0–15 in the quarterly binomial model.

[0] "down, down, down, down" pass through nodes 16, 32, 48, 64 and 1, 17, 33, 49, respectively.

In general, a small, manageable number of scenarios has to be selected or sampled from a large set. To reduce the number of scenarios while keeping reliable results, we applied three different approaches:

- Use of the nonrandom sampling strategy of Zenios and Shtilman (1993). Here, all possible paths over several first periods are fully covered proceeding then with alter-

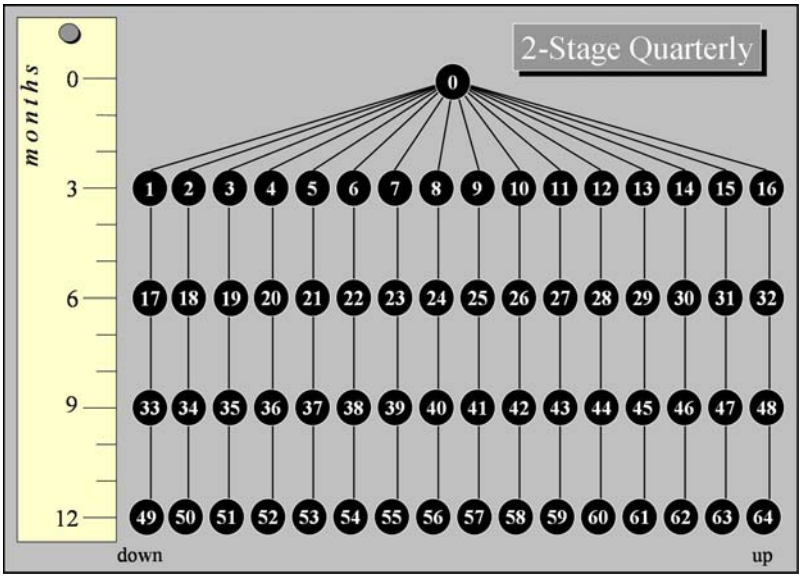


Fig. 3. Fan of scenarios and nodes.

nating up-down movements. For example, when covering fully t periods we get 2^t scenarios; the acronyms are ZS(No. of scenarios);

- Ad hoc selected scenarios and/or scenario trees. For example, one uses a particular subset of scenarios along the planning horizon and proceeds with alternating up-down movements. Figure 4 is the case of selecting 8 particular scenarios from the 2^{12} scenarios for the planning horizon of 12 months; the acronym is Part(8). The "expert-designed" scenario tree may be put into this group: The expert considers a possibility to adapt to information obtained at the end of the first quarter, formulated simply as an increase or a decrease of interest rates. Instead of the multiperiod two-stage program based on the scenarios obtained from the BDT lattice he solves a three stage program based on the same set of scenarios, with the initial known interest rate r_0, two possible descendants "down" or "up" followed each by 8 possible continuations. See Figure 5.
- Use the scenario reduction method GAMS-SCENRED based on results of Dupačová, Gröwe-Kuska and Römisch (2003) and described in Gröwe-Kuska, Heitsch and Römisch (2003). See the next section for a brief introduction.

Except for the initial market price of the bonds, which is observable, the fair price of the jth bond at time t just after the coupon was paid and under scenario s follows for each scenario by formula (1) and formulas like (23) at branching nodes of scenario trees for multistage problems.

	Investment time horizon													
Scenario	1	2	3	4	5	6	7	8	9	10	11	12	13	continue
0	down	down	down	down	down	down	down	down	down	down	down	down	down	up&down
1	down	down	down	up	down	down	down	down	down	up	down	down	down	up&down
2	down	up	down	down	down	up	down	up	down	down	down	up	down	up&down
3	up	down	up	down	up	down	up	down	up	down	up	down	up	up&down
4	up	down	up	down	up	up	down	up	up	down	up	down	up	up&down
5	up	down	up	up	up	down	up	down	up	up	up	down	up	up&down
6	up	up	up	down	up	up	up	up	up	down	up	up	up	up&down
7	up	up	up	up	up	up	up	up	up	up	up	up	up	up&down

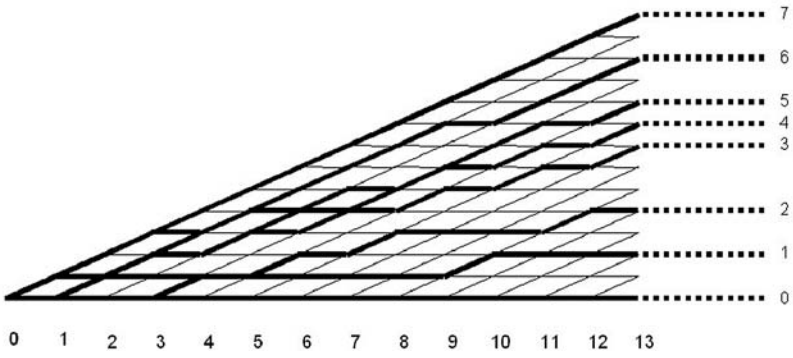


Fig. 4. Part(8) sampling strategy.

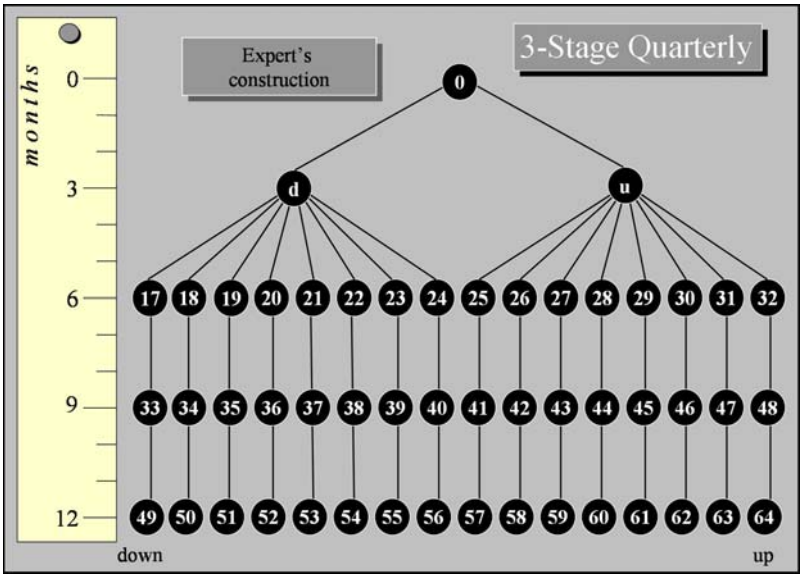


Fig. 5. Expert-designed 3-stage scenario tree.

4. Scenario reduction and scenario tree construction

The idea developed in Dupačová, Gröwe-Kuska and Römisch (2003) and extended in Gröwe-Kuska, Heitsch and Römisch (2003) is a natural one. It is based on quantitative stability results valid for a large class of stochastic programming problems of the form

$$\max_{x\in\mathcal{X}} F(x, P) := \int_{\Omega} f(x, \omega) P(\mathrm{d}\omega). \tag{25}$$

Here, $\mathcal{X} \neq \emptyset$ is a fixed closed, convex set, $\omega \in \Omega \subset R^{\tau}$ is random with probability distribution P, f is convex in $\boldsymbol{x}$ and such that the integral is well defined. This fits our problem formulations, for example, for problem (22), $\mathcal{X} = \mathcal{X}_0$ is defined by nonnegativity constraints and by (5)–(6) and P is a discrete probability distribution carried by equiprobable scenarios obtained from the binomial BDT lattice. One may prefer to approximate P by another probability distribution Q. Quantitative stability results claim, inter alia, that closeness of probability distributions P, Q is important for closeness of optimal values $\varphi(P)$, $\varphi(Q)$ of the two corresponding programs (25). We apply the Kantorovich distance $D_K(P, Q)$ of P, Q which for two *discrete* probability distributions is equal to the optimal value of a transportation problem. The result is then the starting point for designing scenario reduction algorithms which aim at selecting a (small) subset of scenarios and assigning them probabilities so that the distance of the reduced probability distribution Q from the original P is kept small.

With P carried by scenarios ω^i, $i = 1, \ldots, S$, with probabilities p_i and for Q carried by scenarios $\bar{\omega}^j$, $j = 1, \ldots, \bar{S}$, with probabilities q_j, the Kantorovich distance is

$$D_K(P, Q) = \min\Bigg\{ \sum_{i=1}^{S}\sum_{j=1}^{\bar{S}} \eta_{ij} c_{\tau}\big(\omega^i, \bar{\omega}^j\big): \eta_{ij} \geqslant 0,\ \sum_{i=1}^{S} \eta_{ij} = q_j,\ \sum_{j=1}^{\bar{S}} \eta_{ij} = p_i\ \forall i, j \Bigg\} \tag{26}$$

where $c_{\tau}(\omega^i, \bar{\omega}^j)$ is a suitably defined distance between scenarios.

In our problem, scenarios are paths of realizations of the n-dimensional stochastic ($n = 15$) process ω with components the risk-free rate, the prices of the bonds in portfolio and the cashflows over time. The tth realization ω_t of ω is the vector defined by the rate which applies in period $(t, t + 1]$, the prices P_{jt} of bonds in portfolio that are induced by rates through (1) and (23) and the cashflows f_{jt}. Probability distribution P corresponds to all scenarios obtained from the full lattice up to the planning horizon, i.e., for $\tau = T$, whereas probability distribution Q is carried only by a subset of these scenarios. We choose

$$c_T\big(\omega^i, \bar{\omega}^j\big) = \sum_{t=1}^{T-1} \big|\omega_t^i - \bar{\omega}_t^j\big| \tag{27}$$

where $|\cdot|$ denotes some norm in R^n. When deleting only one scenario from ω^i, $i = 1, \ldots, S$, its index i^* is determined to minimize over all $i = 1, \ldots, S$ the product $p_i \min_{i \neq j} c_T(\omega^i, \omega^j)$. The initial probability p_{i^*} of this scenario is then added to the probability of scenario $\omega^{j(i^*)}$ such that the distance $c_T(\omega^{i^*}, \omega^{j(i^*)})$ is minimal among all scenarios $\omega^j \neq \omega^{i^*}$. As a result, after reduction of the fan of scenarios obtained from the fitted binomial lattice, the kept scenarios need not be equiprobable. This idea is used repeatedly in the backward algorithms applied in our study. Similarly, one may develop a simple criterion for the case when only one of scenarios is kept and this is the basic step of the forward algorithm.

It is in principle possible to use any distance in R^{nT} and/or to put positive, decreasing weights to the summands when lower importance of more distant periods rates should be reflected; see, e.g., Pflug (2001).

Various versions of the reduction algorithm provide an "optimally" reduced probability distribution Q^* and also the relative precision obtained. This is measured by the ratio $D_K(P, Q^*)/D_K(P, \delta_{\omega^*})$ where δ_{ω^*} is degenerate distribution carried by one scenario, ω^*, such that its distance from P is minimal, i.e., $D_K(P, \delta_{\omega^*}) = \min_i D_K(P, \delta_{\omega^i})$.

A heuristics for scenario tree construction and an algorithm have been delineated in Gröwe-Kuska, Heitsch and Römisch (2003). The present implementation requires the user to characterize the demanded relative reduction at a given time point t by specifying the number of scenarios to be kept or number of nodes to appear in different levels of the scenario tree, i.e., to specify the stage-wise tree structure. In our numerical experiments we code the desired or prescribed structure of the scenario tree indicating the number of leaves, the number of their predecessors up to number of nodes in the set $\mathcal{N}_1$.

In comparison with clustering techniques, scenario tree construction applies a backward clustering of scenarios into a selected number of clusters at each stage.

5. Numerical results

Table 2 shows the optimal values obtained by solving the two-stage model with monthly time step over one year horizon for the full problem (4096 scenarios) and for the reduced models (with 2048 and 280 scenarios). The results were obtained by the three implemented reduction algorithms available under GAMS v.21.2, named GAMS-SCENRED, cf. Gröwe-Kuska, Heitsch and Römisch (2003):

- method 1 = fast backward–forward algorithm,
- method 2 = fast backward–backward algorithm,
- method 3 = fast backward algorithm.

Table 2 also shows the optimal values obtained by solving the two-stage quarterly model and its three-stage expert-designed variant over the one year horizon for the full problem (16 scenarios) and the reduced one (with 8 scenarios). For both discretizations, the date of October 3, 1994 is considered.

In Table 3, we report the results obtained in Bertocchi, Dupačová and Moriggia (2000) for two-stage models with monthly discretization using the ad hoc chosen sce-

Table 2
Optimal function values for different numbers of scenarios and different time steps

October 3, 1994	No. of scen.	Full	No. scen. reduced	Met. 1	Met. 2	Met. 3
2-stage monthly	4096	11 436.6	2048	11 443.21	11 443.20	11 433.18
2-stage monthly	4096	11 436.6	280	11 443.13	11 443.20	11 433.20
2-stage quarterly	16	11 440.43	8	11 445.42	11 442.73	11 446.55
3-stage quarterly	16	11 439.98	8	11 499.38	11 499.63	11 512.73

Table 3
Optimal function values for reduction and sampling techniques

September 1, 1994	No. of scen.	Full	No. scen. reduced	Met. 1	Met. 2	Met. 3
2-stage monthly	4096	11 697.34	2048	11 704.08	11 704.09	11 704.09
2-stage monthly	4096	11 697.34	280	11 704.08	11 704	11 704.09
ZS(8)	8	11 560	8	11 700.51	11 700.24	11 700.11
ZS(16)	16	11 472	16	11 702.59	11 703.91	11 704.13
ZS(32)	32	11 559	32	11 703.98	11 703.96	11 704.05
ZS(64)	64	11 470	64	11 704.09	11 704.09	11 704.09
Part(8)	8	11 499	–	–	–	–

nario selection strategy (Part(8)) and the Zenios–Shtilman strategy, cf. Zenios and Shtilman (1993), from 8 to 64 equiprobable scenarios. We compare the optimal values with those obtained for the full lattice (4096 scenarios) and reduced lattices (2048, 280, 64, 32, 16, 8 scenarios) resulting by application of the three SCENRED methods where the probabilities of the resulting scenarios are no more equal. The reduction algorithms work well also in the case of a large reduction in the number of scenarios; indeed for our data, the reduction algorithms seem to work better than the applied sampling techniques.

These results refer to the stochastic bond portfolio management problem as formulated in Section 2, without any additional constraints on cash and without limitation on the holding of a specific bond in the portfolio. The optimal values obtained for the reduced models overestimate the sought maximal value of the expected wealth at the horizon for the full problem, see Tables 2 and 3.

We run also the same stochastic bond portfolio management problem with further constraints on cash (less than 1000) and on the amount of puttable bond CTO13212 (less

than 30). This formulation results in a more diversified first-stage solution; however, the behavior in terms of the optimal function value for the full and reduced models does not differ from those reported in Tables 2 and 3.

Below we illustrate the outcomes of the scenario reduction and construction methods for the quarterly discretization and for the planning horizon of one year. This small size example allows for various experiments and helps to discover the overall features of these new techniques.

We start with the quarterly two-stage fan of 16 scenarios, see Figure 3. The scenarios are represented by arcs and nodes, one node for each scenario and quarter.

Figure 7 illustrates the tree that we got by the *scenario reduction approach* using the fast backward algorithm of Dupačová, Gröwe-Kuska and Römisch (2003). The chosen number of leaves is 8 (i.e., 50% reduction). The acronym of this structure is 8-8-8-8 (numbers of nodes at the end of the four quarters). The scenarios containing "white nodes" have been deleted during the reduction.

Figures 8–11 have been obtained by applying the *construction* procedure when prescribing various nonincreasing numbers of nodes at each quarter starting with the last quarter.

We report in Table 4 the optimal value, absolute distance and relative distance for the date of October 3, 1994 for our bond portfolio problem in Bertocchi, Dupačová and Moriggia (2006) for all the cases related to Figures 3 and 7–11. The reduced problem 8-8-8-8, the two-stage problem with 8 scenarios, is the closest to the original one in terms of the Kantorovich distance of the two probability distributions P and Q. We also report parallel results for the modified bond portfolio model where two more constraints were introduced, one limiting the cash in portfolio at the planning horizon and one limiting the quantity of puttable bond in the portfolio.

Table 4
Optimal function values, absolute and relative distance for various tree structures

October 3, 1994	16 scen.	8-8-8-8 scen.	8-8-8-2 scen.	8-6-4-2 scen.	8-4-2-2 scen.	16-8-4-2 scen.
Optimal value without constraints	11 440	11 446	11 716	11 428	11 634	11 622
Optimal value with constraints	11 387	11 389	11 603	11 386	11 547	11 535
Absolute distance	–	0.000952	0.001381	0.002014	0.002227	0.001801
Relative distance	–	0.32	0.46	0.68	0.75	0.61

Table 5
Optimal function values in stress testing on the 8 reduced scenario tree

October 3, 1994	8 scen.	8 + [2] scen.	8 + [5] scen.	8 + [7] scen.	
Case 1 Optimal values	11 512.73	11 512.69	11 470.94	11 472.93	
	8 + [8] scen. 11 516.63	8 + [9] scen. 11 510.17	8 + [10] scen. 11 512.93	8 + [11] scen. 11 511.43	8 + [14] scen. 11 514.08
October 3, 1994	8 scen.	8 + [2] scen.	8 + [5] scen.	8 + [7] scen.	
Case 2 Optimal values	11 512.73	11 536.61	11 462.4	11 430.41	
	8 + [8] scen. 11 545.19	8 + [9] scen. 11 508.16	8 + [10] scen. 11 508.2593	8 + [11] scen. 11 504.33	8 + [14] scen. 11 509.04

The reduction and construction algorithms do not depend on constraints, hence, the same scenarios are selected for the unconstrained case and the constrained one, and the absolute and relative distances remain unchanged, too.

Moreover, we considered again the expert-designed three-stage quarterly model on October 3, 1994, see Figures 5 and 12 for the full scenario tree and the reduced one. The difference between the reduced structures depicted on Figures 7 and 12 comes from the structure of the reduction technique which is based on all the information (rates, prices and cashflows) available at each node. We report in Table 5 results of the what-if analysis when the reduced expert-designed scenario tree (8 scenarios with fast backward method) is extended for an additional scenario. Two ad hoc strategies for redistributing probabilities among scenarios are considered when an additional scenario is inserted. In Case 1, the additional scenario is included with probability $\lambda = \frac{1}{16} = 0.0625$ and the probability of the most probable closest node is decreased for this value. We also show how the optimal solution changes when assigning to the new scenario probability $\lambda = \frac{1}{9} = 0.11$ (Case 2), and decreasing uniformly probabilities of all initial scenarios. See also Figure 13 for an illustration.

These results provide a feeling on the contribution of adding one single scenario to the optimal value, which is a type of a robustness or stress test. However, it is an ad hoc idea which does not give us a general formulation of a suitable approach. Another possibility is discussed in the next section.

We also treated the expert-designed scenario tree, see Figure 5 as if it were composed of two separate subtrees, obtained as realization of "down" value of interest rate and "up" value of interest rate in the first quarter, respectively. Each scenario subtree may be seen as an ad hoc discrete distribution. We applied the construction technique to the

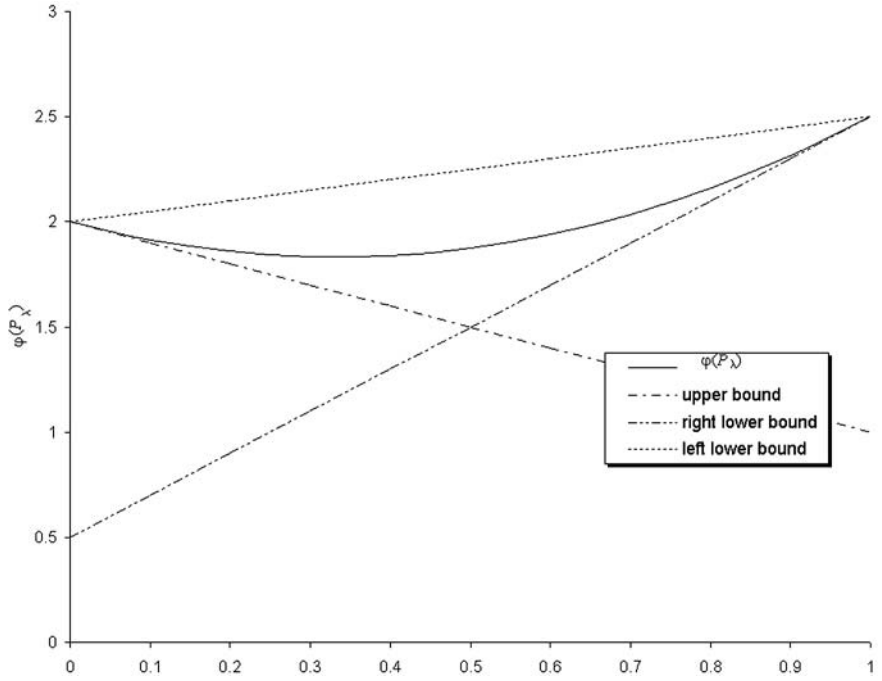


Fig. 6. Upper and lower bounds of the optimal value function.

two separate subtrees: we considered a reduction 5-2-2-1 on the left subtree and 3-2-2-1 (see Figure 14) on the right subtree obtaining the optimal value of 11 431.94 in the former and 11 344.18 in the latter. In the next section we discuss how to combine the distributions on the two subtrees.

6. Stress testing via contamination: Add worst-case scenarios

We assume that the multiperiod or multistage bond portfolio management problem was solved for an already constructed scenario tree; denote again P the corresponding discrete probability distribution, $\varphi(P)$ the optimal value and $x^*(P)$ one of optimal solutions. In some cases, see Figure 9, the extremal scenarios [0], [15] are not fully reflected. The question is to what extent are those scenarios influential with respect to the maximal expected wealth—a problem related with stress testing. An answer may be found by solving the extended problem as we did in experiments summarized in Table 5. For large problems, however, this may be prohibitively expensive. The commonly used stress techniques either evaluate the outcome of the obtained optimal portfolio when these additional scenarios occur or get the optimal value for the probability distribution carried only by the additional scenarios. We use instead the contamination technique to

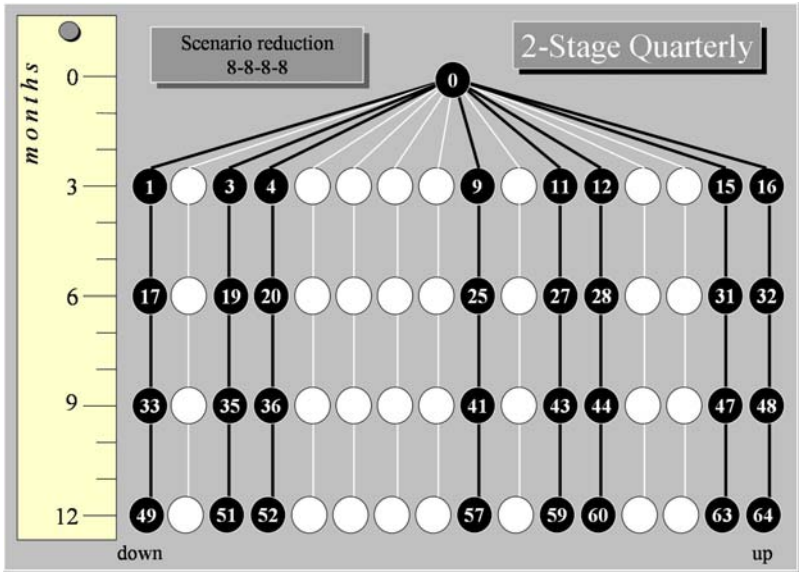


Fig. 7. 2-stage reduced scenario tree: 8-8-8-8 scenarios.

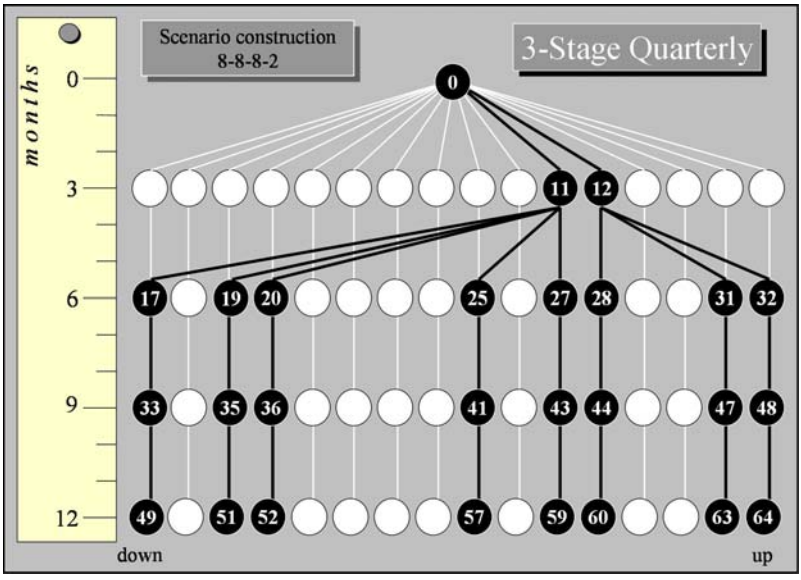


Fig. 8. 3-stage reduced scenario tree: 8-8-8-2 scenarios.

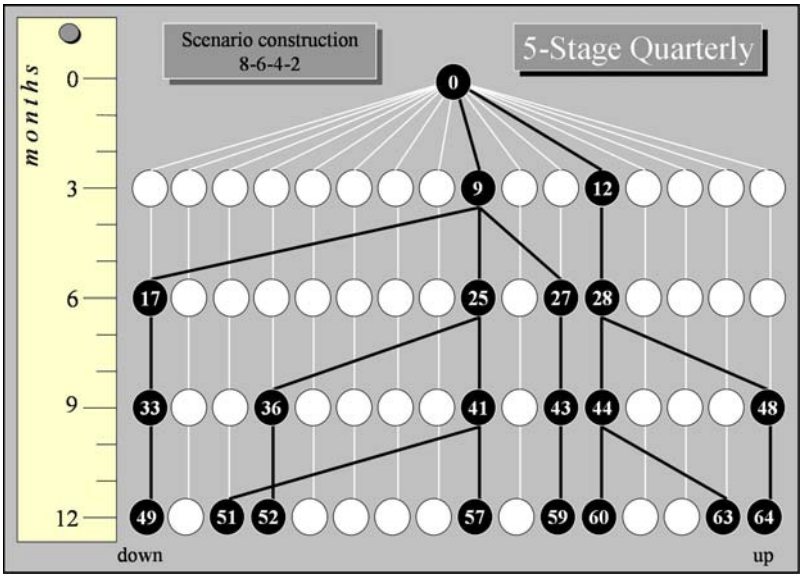


Fig. 9. Reduced scenario tree: 8-6-4-2 scenarios.

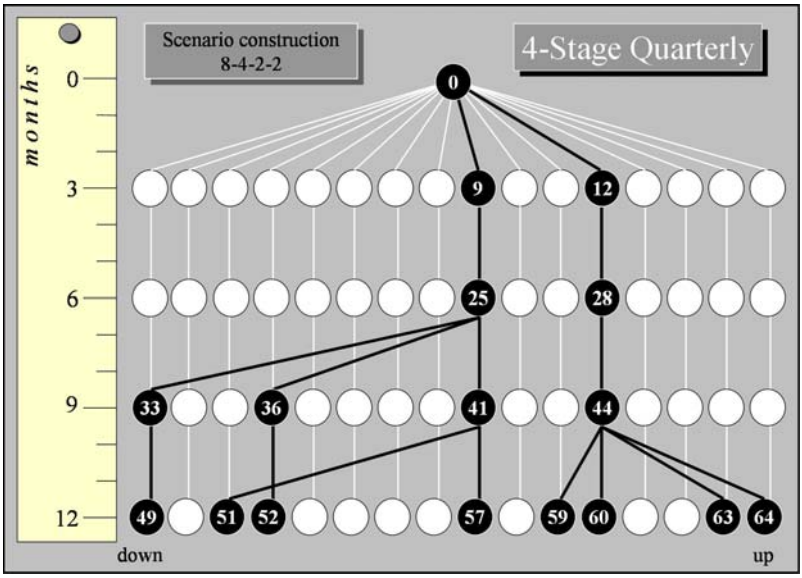


Fig. 10. Reduced scenario tree: 8-4-2-2 scenarios.

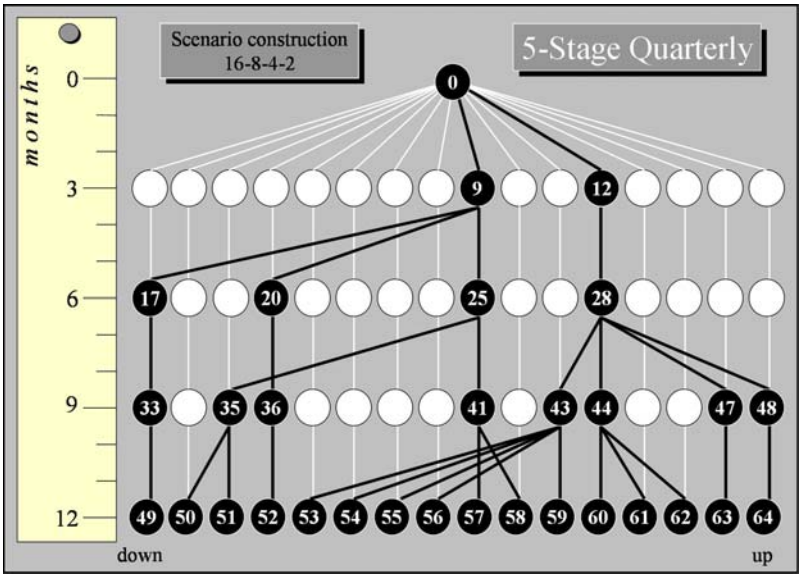


Fig. 11. Reduced scenario tree: 16-8-4-2 scenarios.

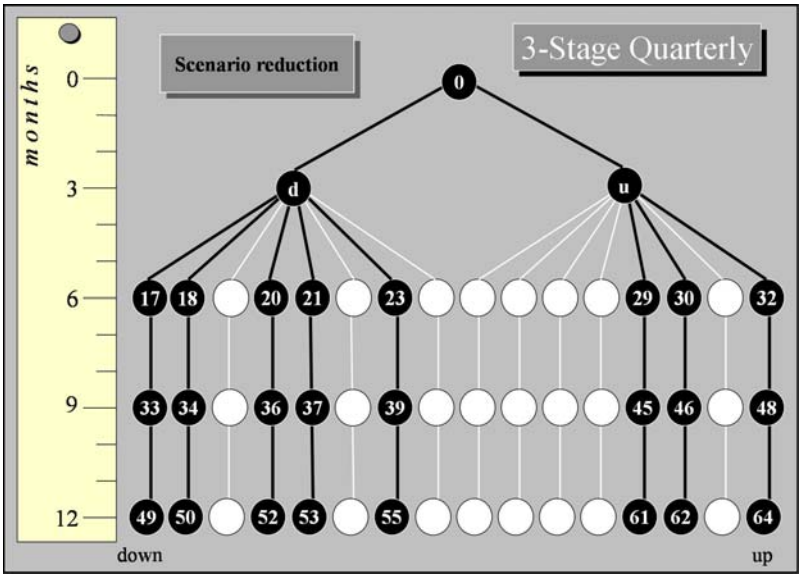


Fig. 12. Expert designed 3-stage reduced scenario tree.

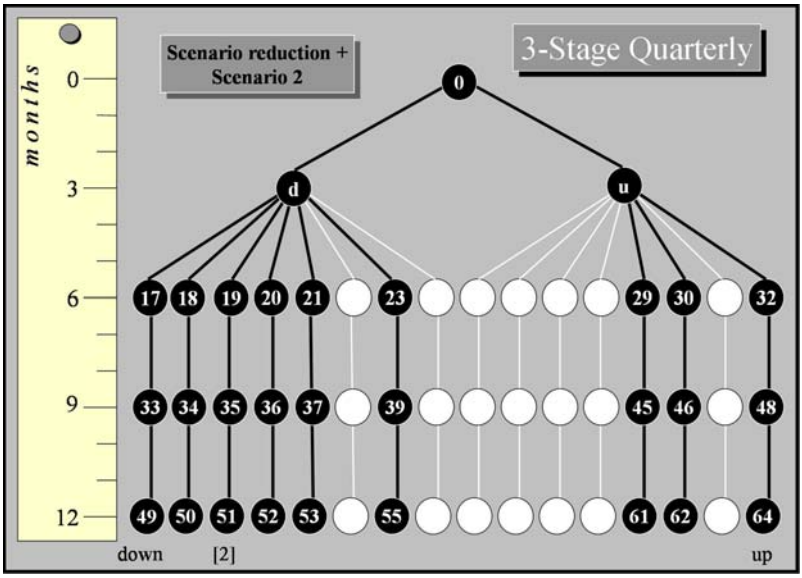


Fig. 13. Expert designed reduced scenario tree with added scenario [2].

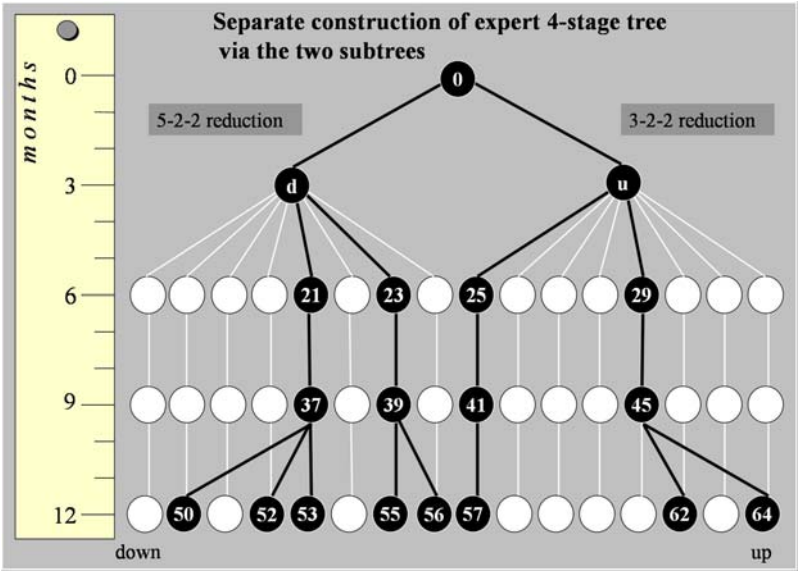


Fig. 14. Expert designed 4-stage reduced scenario subtrees.

quantify the influence of including additional, stress scenarios on the optimal value of the program, i.e., on the maximal expected wealth attained at the planning horizon T. The additional (equiprobable) scenarios, or one of them only, are atoms of the contaminating probability distribution Q. Their inclusion with weight λ means to consider contaminated probability distribution

$$P_\lambda = (1-\lambda)P + \lambda Q \tag{28}$$

carried by the pooled sample of the initial scenarios, atoms of P, and the additional ones, atoms of Q. Then the optimal value $\varphi(P_\lambda)$ of the problem with the extended set of scenarios can be bounded as follows

$$(1-\lambda)\varphi(P) + \lambda\varphi(Q) \geqslant \varphi(P_\lambda) \geqslant (1-\lambda)\varphi(P) + \lambda E_Q W_T\big(x^*(P), \omega\big),$$
$$\lambda \in [0,1]; \tag{29}$$

here $E_Q W_T(x^*(P), \omega) - \varphi(P)$ equals the one-sided, directional derivative of $\varphi(P_\lambda)$ provided that the optimal solution $\boldsymbol{x}^*(P)$ is unique; see Dupačová (1995); Dupačová, Bertocchi and Moriggia (1998).

To apply these contamination bounds, it is necessary to solve the bond portfolio problem twice, not only for P but also for the contaminating distribution Q, and to *evaluate* the expected wealth attained at the horizon for the initial optimal decision $x^*(P)$ when the distribution Q is used instead of P. The contamination bounds take into account also the magnitude of the contamination parameter $\lambda \in (0,1)$. It can be chosen to get equiprobable scenarios in the pooled sample, small values of λ may appear when testing influence of a rare stress scenario, etc.

Similarly, one may think about P_λ as of the distribution Q contaminated by P with contamination weights $\mu = 1-\lambda$ and to construct the bounds in a similar way. Whereas the upper bound in (29) does not change, the lower bound is now based on the expectation $E_P W_T(x^*(Q), \omega)$ where $\boldsymbol{x}^*(Q)$ is an optimal solution of the bond portfolio management problem with distribution Q at the place of P; see Figure 6.

Hence, using the contamination technique we may avoid a complete solution of an extended problem if the obtained bounds, which quantify the influence of including additional scenarios, are tight enough.

In the three-stage problem 8-8-8-2, see Figure 8 and the related short rate scenarios marked $A, \ldots, H$ on Figure 15, contrary to the expert designed 8-8-8-2 problem from Figure 12, the interest rates going "down" in the first quarter were not included. The idea is to check how much their inclusion may change the optimal value. To this end, consider contamination of the probability distribution P carried by the 8-8-8-2 tree by degenerated distributions carried by single scenarios ω^j, $j = [1], [4], [5], [6], [7]$. The question is which of these scenarios may cause the biggest change in the optimal value when included with a small probability λ. As discussed in Dupačová (1995), this is the scenario for which the directional derivative $E_P W_T(x^*(P), \omega^j) - \varphi(P)$ is maximal. In our example, the maximum is attained for $j = [4]$ and the value for scenario $j = [1]$ is almost the same. The results of contamination by these two scenarios are given in

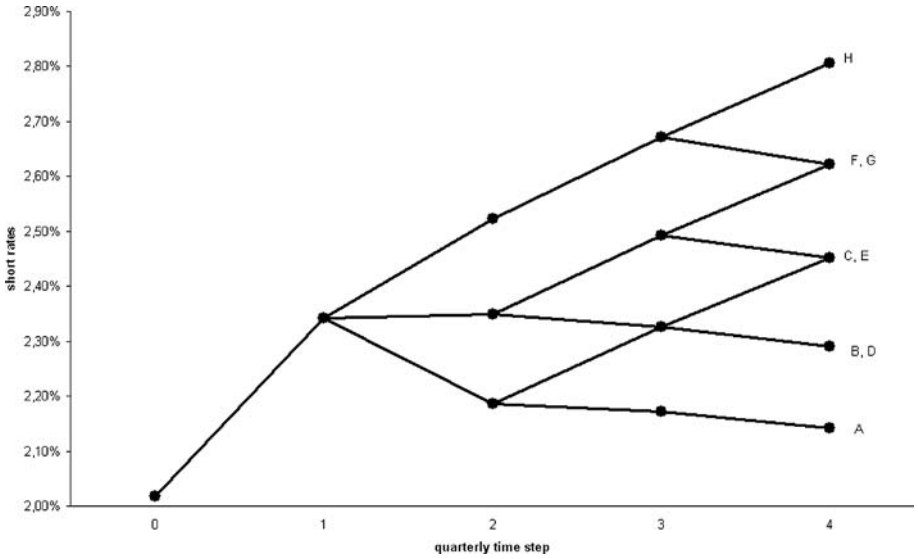
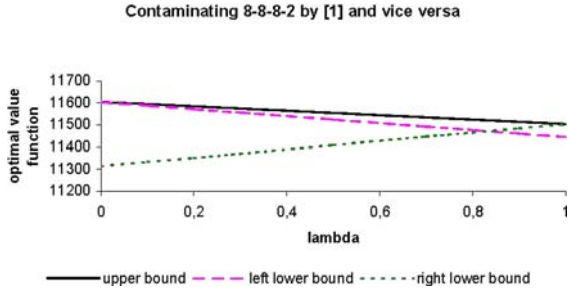


Fig. 15. Short rate scenarios.

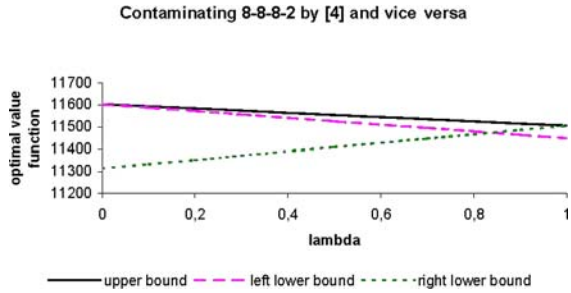
Contaminating 8-8-8-2 by [1] and vice versa

	scenarios set	optimal value function	lambda	upper bound	left lower bound	right lower bound	first derivative contamin.	expectation
P	8-8-8-2	11 603.89	0	11 603.89	11 603.89	11 314.28	−159.3	11444.59
			0.1	11 593.84	11 587.96	11 333.19		
			0.2	11 583.78	11 572.03	11 352.1		
			0.3	11 573.73	11 556.1	11 371		
			0.5	11 553.63	11 524.24	11 408.82		
			0.7	11 533.52	11 492.38	11 446.64		
			0.8	11 523.47	11 476.45	11 465.54		
			0.9	11 513.41	11 460.52	11 484.45		
Q	[1]	11 503.36	1	11 503.36	11 444.59	11 503.36	−189.08	11 314.28



Contaminating 8-8-8-2 by [4] and vice versa

	scenarios set	optimal value function	lambda	upper bound	left lower bound	right lower bound	first derivative contamin.	expectation
P	8-8-8-2	11 603.89	0	11 603.89	11 603.89	11 314.28	−154.03	11 449.86
			0.1	11 594.12	11 588.49	11 333.47		
			0.2	11 584.34	11 573.08	11 352.65		
			0.3	11 574.57	11 557.68	11 371.84		
			0.5	11 555.02	11 526.88	11 410.21		
			0.7	11 535.47	11 496.07	11 448.58		
			0.8	11 525.69	11 480.67	11 467.77		
			0.9	11 515.92	11 465.26	11 486.95		
Q	[4]	11 506.14	1	11 506.14	11 449.86	11 506.14	−191.86	11 314.28



Contaminating 8-8-8-2 by [1,4] and vice versa

	scenarios set	optimal value function	lambda	upper bound	left lower bound	right lower bound	first derivative contamin.	expectation
P	8-8-8-2	11 603.89	0	11 603.89	11 603.89	11 314.28	−156.67	11 447.22
11 588.23			0.1	11 593.98	11 588.22	11 333.33		
11 572.56			0.2	11 584.06	11 572.56	11 352.37		
11 556.89			0.3	11 574.15	11 556.89	11 371.42		
11 525.55			0.5	11 554.32	11 525.56	11 409.52		
11 495.57			0.7	11 534.49	11 494.22	11 447.61		
11 484.95			0.8	11 524.58	11 478.55	11 466.66		
11 487.42			0.9	11 514.66	11 462.89	11 485.7		
Q	[1.4]	11 504.75	1	11 504.75	11 447.22	11 504.75	−190.47	11 314.28

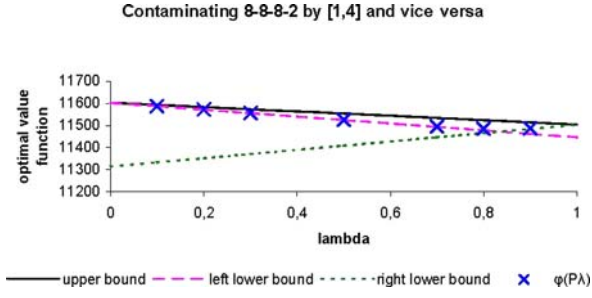


Exhibit 1. Contaminating bounds (case: 8-8-8-2 + [1], [4], [14]).

Contaminating 5-2-2 by 3-2-2 and vice versa

	scenarios set	optimal value function	lambda	upper bound	left lower bound	right lower bound	first derivative contamin.	expectation
P	5-2-2	11 431.94	0	11 431.94	11 431.94	11 421.35	−88.31	11 343.63
11 423.11			0.1	11 423.16	11 423.11	11 413.63		
11 414.28			0.2	11 414.39	11 414.28	11 405.92		
11 405.45			0.3	11 405.61	11 405.45	11 398.2		
11 387.78			0.5	11 388.06	11 387.79	11 382.77		
11 370.11			0.7	11 370.51	11 370.12	11 367.33		
11 361.28			0.8	11 361.73	11 361.29	11 359.61		
11 352.45			0.9	11 352.96	11 352.46	11 351.9		
Q	3-2-2	11 344.18	1	11 344.18	11 343.63	11 344.18	77.17	11 421.35

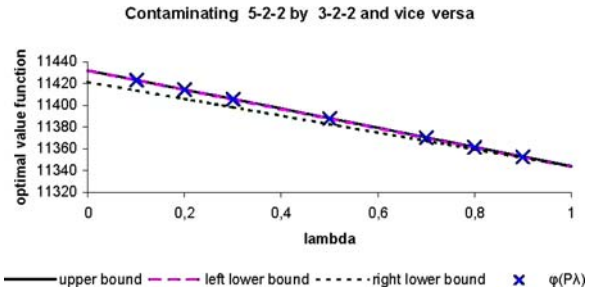


Exhibit 2. Contaminating bounds (case: 5-2-2 vs. 3-2-2).

Exhibit 1 along with contamination by the distribution carried by these two scenarios each with probability 0.5.

The reported results are related to the problem with the additional constraints (cash and puttable bond's limitation). For comparison, optimal values obtained by solution of this problem when contaminated by equiprobable scenarios [1], [4] are marked for several values of λ.

In Exhibit 2 we show the results of contaminating distribution P carried by scenarios on the left subtree of Figure 14 by distribution Q carried by scenarios on the right subtree of the same figure. By choosing the value λ, the expert may express his expectations concerning market developments in the coming quarter, value $\lambda = 0.5$ is induced by the BDT lattice. For comparison, the corresponding maximal expected values of wealth for the combined tree were computed for several values of λ and are marked by crosses.

Finally, in Exhibit 3 we show the contamination of the expert designed 3-stage scenario tree in Figure 5 by an ad hoc constructed scenario with all short rates equal to the initial short rate r_0.

In this way, contamination points out at the out-of-sample scenarios or scenario trees who may lead to the most significant changes of the optimal value, i.e., of the maximal attainable expected wealth at the horizon of the bond portfolio management problem. The additional scenarios need not relate to the BDT model we use for generation of scenarios. Expert designed stress or even catastrophic scenarios may be tested as well.

Contaminating expert full by one ad-hoc and vice versa

	scenarios set	optimal value function	lambda	upper bound	left lower bound	right lower bound	first derivative contamin.	expectation
P	expert full	11 439.98	0	11 439.98	11 439.98	10 716.1	−112.76	11 327.22
			0.1	11 667.12	11 428.7	11 015.62		
			0.2	11 894.25	11 417.43	11 315.15		
			0.3	12 121.39	11 406.15	11 614.67		
			0.5	12 575.66	11 383.6	12 213.72		
			0.7	13 029.93	11 361.05	12 812.76		
			0.8	13 257.06	11 349.77	13 112.28		
			0.9	13 484.2	11 338.5	13 411.81		
Q	one ad-hoc	13 711.33	1	13 711.33	11 327.22	13 711.33	−2995.23	10 716.1

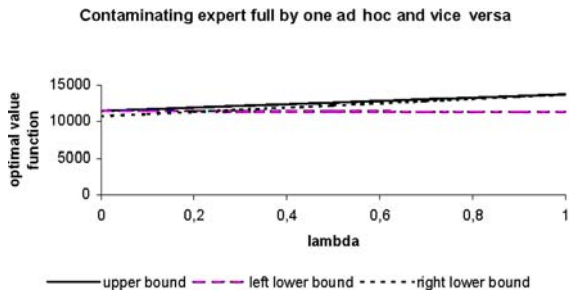


Exhibit 3. Contaminating bounds (case: full-lattice vs. ad hoc scenario).

7. Conclusions

Applications of stochastic programming models in portfolio management give an evidence that this is a powerful and viable approach which is flexible enough to deal with various constraints and is able to capture dynamic aspects of the real-life problems. However, it is not always easy to decide which is the most relevant input structure and/or the form of the scenario-based portfolio management model. We described how a sensible input may be generated and analyzed the results in dependence on the chosen time discretization and the structure of the model. The techniques of Dupačová, Gröwe-Kuska and Römisch (2003); Gröwe-Kuska, Heitsch and Römisch (2003) help us to understand the influence of different model design. The contamination technique, cf. Dupačová (1995), is useful in stress testing with respect to inclusion of additional or stress scenarios. It also helps to identify the most influential among the deleted scenarios.

Even if explained and illustrated here on a simple bond portfolio management problem with interest rate scenarios generated by the Black–Derman–Toy model, the introduced techniques are of general use. They may be applied to more complicated asset-liability management models and for scenarios generated by other methods.

Acknowledgements

The authors acknowledge the support given by research projects MIUR 40% 2002, MIUR 60% 2002 "Methods to analyse bankruptcy risk", research project MSM 0021620839 "Methods of modern mathematics and their applications" financed by MSMT and by Grant Agency (grants 201/05/2340, 402/05/0115) of the Czech Republic.

References

Bertocchi, M., Dupačová, J., Moriggia, V., 2000. Sensitivity of bond portfolio's behavior with respect to random movements in yield curve: A simulation study. Annals of Operations Research 99, 267–286.

Bertocchi, M., Dupačová, J., Moriggia, V., 2006. Horizon and stages in applications of stochastic programming in finance. Annals of Operations Research 142, 67–78.

Bertocchi, M., Giacometti, R., Zenios, S.A., 2005. Risk factor analysis and portfolio immunization in the corporate bond market. European Journal of Operational Research 161, 348–363.

Birge, J.R., Louveaux, F., 1997. Introduction to Stochastic Programming. Springer-Verlag, New York.

Black, F., Derman, E., Toy, W., 1990. A one-factor model of interest rates and its application to treasury bond options. Financial Analysts Journal, 33–39, January/February.

Bradley, S.P., Crane, D.B., 1972. A dynamic model for bond portfolio management. Management Science 19, 139–151.

Cariño, D.R., Turner, A.L., 1998. Multiperiod asset allocation with derivative assets. In: Ziemba, W.T., Mulvey, J.M. (Eds.), World Wide Asset and Liability Modeling. Cambridge University Press, Cambridge, England, pp. 182–204.

de La Grandville, O., 2006. Protecting investors against changes in interest rates: some new results, in this volume.

Dembo, R.S., 1993. Scenario immunization. In: Zenios, S.A. (Ed.), Financial Optimization. Cambridge University Press, Cambridge, England, pp. 290–308.

Dupačová, J., 1995. Postoptimality for multistage stochastic linear programs. Annals of Operations Research 56, 65–78.

Dupačová, J., 2000. Stability properties of a bond portfolio management problem. Annals of Operations Research 99, 267–286.

Dupačová, J., Bertocchi, M., 2001. From data to model and back to data: A bond portfolio management problem. European Journal of Operations Research 134, 261–278.

Dupačová, J., Bertocchi, M., Moriggia, V., 1998. Postoptimality for scenario based financial models with an application to bond portfolio management. In: Ziemba, W.T., Mulvey, J. (Eds.), World Wide Asset and Liability Modeling. Cambridge University Press, Cambridge, England, pp. 263–285.

Dupačová, J., Consigli, G., Wallace, S.W., 2000. Scenarios for multistage stochastic programs. Annals of Operations Research 100, 25–53.

Dupačová, J., Gröwe-Kuska, N., Römisch, W., 2003. Scenario reduction in stochastic programming: An approach using probability metrics. Mathematical Programming 95, 493–511.

Dupačová, J., Hurt, J., Štěpán, J., 2002. Stochastic Modeling in Economics and Finance. Kluwer Acad. Publ., Dordrecht.

Frauendorfer, K., Marohn, Ch., 1998. Refinement issues in stochastic multistage linear programming. In: Marti, K., Kall, P. (Eds.), Stochastic Programming Methods and Technical Applications. In: Lecture Notes in Economics and Mathematical Systems, vol. 458. Springer-Verlag, Berlin, pp. 305–328.

Golub, B., Holmer, M., McKendall, R., Pohlman, L., Zenios, S.A., 1995. A stochastic programming model for money management. European Journal of Operations Research 85 (2), 282–296.

Gröwe-Kuska, N., Heitsch, H., Römisch, W., 2003. Scenario reduction and scenario tree construction for power management problems. In: Advanced Risk Management Using Stochastic Programming with Financial Applications, GAMS Workshop, Heidelberg, September 1–3.

Hiller, R.S., Eckstein, J., 1994. Stochastic dedication: Designing fixed income portfolios using massively parallel Benders decomposition. Management Science 39, 1422–1438.

Hiller, R.S., Schaack, C., 1990. A classification of structured bond portfolio modeling techniques. The Journal of Portfolio Management, 37–48, Fall.

Jamshidian, F., Zhu, Y., 1997. Scenario simulation: theory and methodology. Finance and Stochastics 1, 43–67.

Kahn, R.N., 1991. Fixed income risk modelling. In: Fabozzi, F., et al. (Eds.), The Handbook of Fixed Income Securities, third ed. Business One Irwin, New York, pp. 1307–1319.

Kall, P., Wallace, S.W., 1994. Stochastic Programming. Wiley, Chichester.

Kang, P., Zenios, S.A., 1992. Complete prepayments models for mortgage backed securities. Management Science 38, 1661–1685.

Kouwenberg, R.R.P., Zenios, S.A., 2006. Stochastic programming models for asset liability management, in this volume.

Kusy, M.I., Ziemba, W.T., 1986. A bank asset and liability management model. Operations Research 34, 356–376.

Messina, E., Mitra, G., 1996. Modelling and analysis of multistage stochastic programming problems: A software environment. European Journal of Operations Research 101, 343–359.

Mulvey, J.M., Thorlacius, E., 1998. The Towers Perrin global capital market scenario generation system. In: Ziemba, W.T., Mulvey, J.M. (Eds.), World Wide Asset and Liability Modelling. Cambridge University Press, Cambridge, England, pp. 286–312.

Mulvey, J.M., Zenios, S.A., 1994. Capturing correlations of fixed-income instruments. Management Science 40 (1), 1329–1342.

Nielsen, S.S., Ronn, E.I., 1997. The valuation of default risk in corporate bonds and interest rate swaps. Advances in Futures and Options Research 9, 175–196.

Pflug, G.Ch., 2001. Scenario tree generation for multiperiod financial optimization by optimal discretization. Mathematical Programming 89, 251–271.

Prékopa, A., 1995. Stochastic Programming. Kluwer, Dordrecht and Académiai Kiadó, Budapest.

Shapiro, J.F., 1988. Stochastic programming models for dedicated portfolio selection. In: Mitra, B. (Ed.), Mathematical Models for Decision Support. In: NATO ASI Series, vol. F48. Springer-Verlag, Berlin, pp. 587–611.

Vassiadou-Zeniou, C., Zenios, S.A., 1996. Robust optimization models for managing callable bond portfolios. European Journal of Operational Research 91, 264–273.

Wallace, S.W., Ziemba, W.T. (Eds.), 2005. Applied Stochastic Programming. SIAM – Mathematical Programming Society.

Zenios, S.A. (Ed.), 1993. Financial Optimization. Cambridge University Press, Cambridge, England.

Zenios, S.A., Shtilman, M.S., 1993. Constructing optimal samples from a binomial lattice. Journal of Information & Optimization Sciences 14, 125–147.

Chapter 8

PERTURBATION METHODS FOR DYNAMIC PORTFOLIO ALLOCATION PROBLEMS

GEORGE CHACKO*

Harvard Business School & IFL, USA
e-mail: gchacko@hbs.edu

KARL NEUMAR

John F. Kennedy School of Government, Harvard University & Kite Partners, USA

Contents

Abstract 338
1. Introduction 339
2. General problem formulation 340
   2.1. Investment opportunity set 341
   2.2. Utility function 344
3. Exact solution for unit elasticity of intertemporal substitution 345
   3.1. General results 345
   3.2. Example 1: Time-varying expected returns (finite horizon) 349
   3.3. Example 2: Time-varying expected returns (infinite horizon) 352
4. Approximate solution for general elasticity of intertemporal substitution 353
   4.1. Perturbation around unit elasticity of substitution 353
   4.2. Perturbation around mean of consumption/wealth ratio 360
5. Example 366
   5.1. Time-varying volatility 367
      5.1.1. Exact solution when $\psi = 1$ 368
      5.1.2. Approximation methods when $\psi \neq 1$ 369
      5.1.3. Numerical results 371
      5.1.4. Limiting behavior of the approximated results 372
      5.1.5. Exact solution for terminal utility of wealth 376
   5.2. Time-varying interest rates 378
      5.2.1. Exact solution when $\psi = 1$ 379
      5.2.2. Approximation methods when $\psi \neq 1$ 380

* Contact author: George Chacko, Harvard University, Graduate School of Business Administration, Boston, MA 02163, USA. Tel. +1 617-495-6884.

Handbook of Asset and Liability Management, Volume 1
Edited by S.A. Zenios and W.T. Ziemba
© 2006 Published by Elsevier B.V.
DOI: 10.1016/S1872-0978(06)01008-8

6. Conclusions 382
References 383

Abstract

This chapter presents two perturbation methods for solving dynamic portfolio allocation/consumption problems under general preferences and a time-varying investment opportunity set. The methods work under general recursive utility and any affine investment opportunity set. The methods offer approximate closed-form solutions. We demonstrate the methods with examples involving time-varying expected returns, volatility, and interest rates.

1. Introduction

It is widely recognized that dynamic consumption/investment problems under uncertainty are extremely difficult to solve. However, in finance this is precisely the type of problem that is routinely encountered when researchers attempt to model decisions made by individuals and institutions. This may be one reason why progress in this area has occurred at a slow pace since the path-breaking papers of Merton (1969, 1971, 1973). For this reason, we offer two perturbation methods that allow for the quick calculation of closed-form solutions to a fairly general consumption/investment problem.

By the Principle of Optimality, the solution to stochastic control problems can be found by solving a partial differential equation (PDE) known as the Bellman equation. The main trouble is that this PDE is usually nonlinear and unsolvable, except for the simple cases of log utility or power utility with constant investment opportunities. If complexity in the form of preferences, time-varying investment opportunities, etc., are introduced, numerical methods offer the only way for generating solutions. However, as the dimensionality of the problem increases, the computational time required of conventional numerical methods increase geometrically. Therefore, numerical methods are usually prohibitive as well.

We offer two perturbation methods for solving the dynamic consumption/investment problem under general preferences and a time-varying investment opportunity set. The restrictions that we impose are that preferences be captured by a recursive utility formulation and that the time-varying investment opportunity set be characterized by an affine differential generator. Under these restrictions, we show that approximate closed-form solutions can be found by utilizing two different perturbation methods, where the perturbation is taken around two exact solutions.

The first perturbation method is essentially a perturbation around the exact solution when the elasticity of substitution parameter in the recursive utility function takes on a value of 1 (and the risk aversion parameter takes on any value). Under this condition, an exact solution can be calculated for the class of dynamic consumption/investment problems with an affine investment opportunity set. We then build on this result to show that an approximate solution to the general problem with arbitrary elasticity of substitution parameter can be constructed by taking a perturbation around this closed-form solution for unit elasticity of substitution.

The second perturbation method is a perturbation around the mean consumption to wealth ratio of the solution.[1] We show that the solution to the general problem with an affine investment opportunity set has the unconditional mean of the consumption–wealth ratio in the Bellman equation. In fact, we show that this one term that creates a nonlinearity in the Bellman equation that makes closed-form solutions intractable. Therefore, one approach to solving this problem is to simply substitute the unconditional mean of the consumption–wealth ratio in place of the actual consumption–wealth

[1] This method has been used for specific models in a number of papers including Campbell (1993), Campbell and Viceira (1999), Chacko and Viceira (2003, 2005) and Viceira (2001).

ratio. This leads to a closed-form solution around which a perturbation can be taken to reflect the difference between the realized consumption–wealth ratio and its unconditional mean. The perturbation leads to an approximate analytic solution to the general consumption-investment problem.

This chapter is organized as follows. In Section 2, we define the generalized problem. Here we provide restrictions on the utility function and the investment opportunity set. One important result we show here is that for general recursive preferences, the time variation in the investment opportunity set can be completely characterized by the Sharpe ratios of the investment opportunities.[2] In Section 3, we present an exact solution to the general problem in the case of unit elasticity of substitution (when elasticity of substitution equals one). We then show how to use this exact solution to derive an approximate solution to non-unit elasticities of substitution. This is presented in Section 4, where we take a perturbation around the unit elasticity of substitution solution. In addition this section also presents a perturbation around the mean consumption–wealth ratio. Section 5 provides further examples, and Section 6 concludes.

2. General problem formulation

Consider the general consumption/investment problem

$$\sup_{\pi,C} E_0\left[\int_0^T u(C_t)\,\mathrm{d}t + Q(W_T)\right], \tag{1}$$

$$\text{s.t.}\quad \mathrm{d}W_t = \left[\pi_t m_t W_t + (1-\pi_t) r_t W_t - C_t\right]\mathrm{d}t + \pi_t W_t \sqrt{v_t}\,\mathrm{d}Z_{S,t}, \tag{2}$$

$$\frac{\mathrm{d}S_t}{S_t} = m_t\,\mathrm{d}t + \sqrt{v_t}\,\mathrm{d}Z_{S,t}, \tag{3}$$

$$\frac{\mathrm{d}B_t}{B_t} = r_t\,\mathrm{d}t \tag{4}$$

where W_t represents the agent's wealth, $Q(W_T)$ represents utility from bequeathing wealth, and $u(C_t, t)$ denotes the instantaneous utility the agent derives from consumption.[3] The instantaneous utility function is assumed to satisfy standard properties (monotonicity and concavity). S_t represents the price of a risky asset, while B_t represents the value of a risk-free asset. The agent chooses how to allocate his wealth between consumption and savings (C_t denotes his consumption decision), and further chooses how to allocate his savings between a risky and risk-free asset (π_t denotes his portfolio

[2] This extends the results in Nielsen and Vassalou (2000, 2004), who also show that the Sharpe ratio is a sufficient statistic but do so only for the case of maximizing terminal utility of wealth.

[3] The special case where $u(C_t) = 0$ and $Q(W_T) = \log W_T$ has been studied extensively; the solution yields the growth-optimal portfolio (MacLean and Ziemba (1991, 1999) and MacLean, Ziemba and Li (2005)). See Hakansson and Ziemba (1995) for thorough review.

allocation decision). The variables m_t, v_t, and r_t denote the instantaneous expected return of the stock, the instantaneous variance of the stock, and the instantaneous interest rate, respectively. These three variables could be time-varying.

2.1. Investment opportunity set

We specify in this section the dynamics of the investment opportunity set. For this chapter, we only consider opportunity sets that are *affine* systems.[4]

The investment opportunity set is assumed to be time-varying through the drifts and diffusion terms of the risky and riskless assets. We assume that the drifts and squared diffusions of each state variable are linear in that state variable so that the state variables and log stock returns form the affine system

$$\begin{aligned} \mathrm{d}m_t &= \mu_m(m_t)\,\mathrm{d}t + \sigma_m(m_t)\,\mathrm{d}Z_{m,t}, \\ \mathrm{d}\left(\frac{1}{v_t}\right) &= \mu_v(v_t)\,\mathrm{d}t + \sigma_v(v_t)\,\mathrm{d}Z_{v,t}, \\ \mathrm{d}r_t &= \mu_r(r_t)\,\mathrm{d}t + \sigma_r(r_t)\,\mathrm{d}Z_{r,t}. \end{aligned} \tag{5}$$

Rather than considering the instantaneous variance as a state variable, we instead consider its reciprocal, the instantaneous precision, as a state variable.

Each state variable is assumed to have a constant instantaneous correlation with stock returns given by $\rho_{\cdot,S}$, e.g., the innovations to expected returns, $Z_{m,t}$, has a correlation of ρ_{mS} with stock return innovations. For the sake of clarity and notational ease, we assume for the remainder of the chapter that the factors are orthogonal to each other. However, it is conceptually simple to extend the results of this chapter to the case of a more general covariance matrix for the factors and stock returns, conditional on the system of state variables remaining affine.

By the Principle of Optimality, the optimal controls to the consumption/investment problem also satisfy the optimization problem

$$\begin{aligned} 0 = \sup_{\pi,C}\Big\{ & u(C) + \big[\pi(m-r)W + rW - C\big]J_W + \frac{1}{2}\pi^2 W^2 v J_{WW} \\ & + \mu_m J_m + \frac{1}{2}\sigma_m^2 J_{mm} + \mu_v J_v + \frac{1}{2}\sigma_v^2 J_{vv} + \mu_r J_r + \frac{1}{2}\sigma_r^2 J_{rr} \\ & + \pi\rho_{mS}\sqrt{v}\sigma_m W J_{mW} + \pi\rho_{vS}\sqrt{v}\sigma_v W J_{vW} + \pi\rho_{rS}\sqrt{v}\sigma_r W J_{rW} + J_t \Big\} \end{aligned} \tag{6}$$

where J represents the indirect utility function and the subscripts on J represent partial derivatives.[5] To simplify notation, we rewrite this as

[4] See Duffie and Kan (1996) and Dai and Singleton (2000) for a comprehensive discussion of affine models in the context of the term structure of interest rates.

[5] The time subscripts have been left off to avoid confusion.

$$0 = \sup_{\pi,C}\Bigg\{u(C) + \big[\pi(m-r)W + rW - C\big]J_W + \frac{1}{2}\pi^2 W^2 v J_{WW} + \sum_{i=m,r,y}\mu_i J_i + \sum_{i=m,r,y}\frac{1}{2}\sigma_i^2 J_{ii} + \sum_{i=m,r,y}\pi\rho_{iS}\sqrt{v}\,\sigma_i W J_{iW} + J_t\Bigg\}$$

where $y \equiv 1/v$.

Under certain circumstances the time variation in the investment opportunity set due to stochastic expected returns, m_t, and/or stochastic volatility, $\sqrt{v_t}$, can be captured by a single state variable, the Sharpe ratio $(m_t - r_t)/\sqrt{v_t} = X_t$. Thus, this section extends Kim and Omberg (1996) who discovered this result in a model with no intermediate consumption and constant interest rates. We show precisely the restrictions that need to hold for this result to apply generally and show the economy in Kim and Omberg (1996) to be a special case.

The following proposition generalizes Kim and Omberg (1996) to show that the Sharpe ratio is generally a sufficient statistic to capture time variation in expected returns and volatility. Assume that the dynamics of the Sharpe ratio is given by

$$\mathrm{d}X_t = \mu_X(r_t, m_t, v_t, X_t)\,\mathrm{d}t + \sigma_X(r_t, m_t, v_t, X_t)\,\mathrm{d}Z_{X,t}. \tag{7}$$

The instantaneous correlation between X_t and r_t are given by a constant ρ_{Xr}. The Sharpe ratio and interest rate also have constant instantaneous correlations with the stock price of ρ_{XS} and ρ_{rS}, respectively.

Proposition 1. *The dynamics of the statistic $X_t = (m_t - r_t)/\sqrt{v_t}$, the Sharpe ratio, is sufficient to capture all of the time variation in the investment opportunity set due to stochastic expected returns and volatility only if the drift and diffusion of the Sharpe ratio, $\mu(r_t, m_t, v_t, X_t)$ and $\sigma_X(r_t, m_t, v_t, X_t)$, can be written as functions of only the interest rate and the Sharpe ratio, i.e., $\mu(r_t, X_t)$ and $\sigma_X(r_t, X_t)$.*

Proof. We prove this by conjecturing that the Sharpe ratio is in fact a sufficient statistic and deriving the Bellman equation. We then verify this conjecture by showing that the resulting Bellman equation does not contain the state variables m_t and v_t, indicating that the optimal consumption and portfolio choice also will not contain these state variables.

By the Principle of Optimality, the optimal portfolio allocation satisfies

$$0 = \sup_{\pi}\Bigg\{u(C) + J_t + \big[\pi(m-r)W + rW - C\big]J_W + \frac{1}{2}\pi^2 v W^2 J_{WW} + \mu_X J_X + \frac{1}{2}\sigma_X^2 J_{XX} + \rho_{XS}\pi\sigma_X\sqrt{v}W J_{WX} + \mu_r J_r + \frac{1}{2}\sigma_r^2 J_{rr} + \rho_{rS}\pi\sigma_r\sqrt{v}W J_{Wr} + \rho_{Xr}\sigma_X\sigma_r J_{Xr}\Bigg\}$$

where J represents the indirect utility function.[6] This equation contains the state variables m_t and v_t. We will show now that these variables drop out of the resulting optimal controls. The first order conditions for optimal consumption and portfolio choice are

$$C = u_C^{-1}(J_W),$$
$$\pi = -\frac{(m-r)J_W}{vWJ_{WW}} - \frac{\rho_{XS}\sigma_X J_{WX}}{\sqrt{v}WJ_{WW}} - \frac{\rho_{rS}\sigma_r J_{Wr}}{\sqrt{v}WJ_{WW}}.$$

Substituting this first order condition into the static optimization equation results in the Bellman equation

$$\begin{aligned} 0 = {} & J_t - \frac{(m-r)^2 J_W^2}{2vJ_{WW}} - \frac{(m-r)\rho_{XS}\sigma_X J_W J_{WX}}{\sqrt{v}J_{WW}} - \frac{(m-r)\rho_{rS}\sigma_r J_W J_{Wr}}{\sqrt{v}J_{WW}} \\ & + rWJ_W - \frac{\rho_{XS}^2\sigma_X^2 J_{WX}^2}{2J_{WW}} - \frac{\rho_{rS}^2\sigma_r^2 J_{Wr}^2}{2J_{WW}} + \mu_X J_X + \frac{1}{2}\sigma_X^2 J_{XX} + \mu_r J_r \\ & + \frac{1}{2}\sigma_r^2 J_{rr} + \rho_{Xr}\sigma_X\sigma_r J_{Xr}. \end{aligned}$$

Replacing all of the instances of $(m-r)/\sqrt{v}$ with the the Sharpe ratio X in the Bellman equation gives

$$\begin{aligned} 0 = {} & u(u_C^{-1}(J_W)) + J_t - \frac{X^2 J_W^2}{2J_{WW}} - \frac{\rho\sigma_X X J_W J_{WX}}{J_{WW}} - \frac{\rho_{rS}\sigma_r X J_W J_{Wr}}{J_{WW}} + rWJ_W \\ & - \frac{\rho^2\sigma_X^2 J_{WX}^2}{2J_{WW}} - \frac{\rho_{rS}^2\sigma_r^2 J_{Wr}^2}{2J_{WW}} + \mu_X J_X + \frac{1}{2}\sigma_X^2 J_{XX} + \mu_r J_r + \frac{1}{2}\sigma_r^2 J_{rr} \\ & + \rho_{Xr}\sigma_X\sigma_r J_{Xr}. \end{aligned}$$

By inspection, observe that the Bellman equation contains only the Sharpe ratio and interest rate as state variables if and only if μ_X and σ_X, the instantaneous expected change and the instantaneous standard deviation of the Sharpe ratio, are functions of only the Sharpe ratio and the interest rate. Therefore, if μ_X and σ_X can be written as a function only of the Sharpe ratio and the interest rate, then the indirect utility function, J, will only be a function of these two state variables. Thus, the Sharpe ratio's dynamics in this case will be a sufficient statistic to capture variation in future investment opportunities related to expected returns and volatility. □

Kim and Omberg (1996) consider a special case of this proposition. In their model, the interest rate is constant and there is no intermediate consumption. More importantly, they use an Ornstein–Uhlenbeck process for the Sharpe ratio. Under this stochastic process, σ_X is a constant, which satisfies the conditions of the proposition above.

However, it is important to realize that the conditions of the proposition are restrictive. Not all stochastic processes for the expected return and volatility will satisfy the

[6] The time subscripts in this equation have been dropped and the subscripts on J denote partial derivatives.

conditions. For example, if the expected return and variance are both uncorrelated diffusions

$$\mathrm{d}m_t = \sigma\,\mathrm{d}Z_{m,t}, \qquad \mathrm{d}v_t = \delta v_t\,\mathrm{d}Z_{v,t},$$

and the interest rate is constant, the conditions are not satisfied. In this example if the interest rate is zero, the dynamics of the Sharpe ratio are given by

$$\mathrm{d}X_t = \delta^2 X_t\,\mathrm{d}t + \frac{\sigma}{v_t}\,\mathrm{d}Z_{m,t} + \delta X_t\,\mathrm{d}Z_{v,t}$$

which does not satisfy the conditions of the proposition because of the $1/v_t$ term in the volatility of the Sharpe ratio.

However, if either m_t or v_t is constant, then the conditions of the proposition will always be satisfied as the Sharpe ratio then represents a simple transformation of a single state variable into another state variable.

Because not all specifications for the investment opportunity set satisfy the conditions of the proposition, we will use the stochastic processes given by (5) as the state variables rather than the Sharpe ratio and the interest rate. This will allow us to consider a much wider class of models.

2.2. *Utility function*

Investor's preferences are described by a recursive utility function. Recursive utility is a generalization of the standard, time-separable power utility model that separates relative risk aversion from the elasticity of intertemporal substitution of consumption.[7] Epstein and Zin (1989, 1991) derive a parameterization of recursive utility in a discrete-time setting, while Duffie and Epstein (1992a, 1992b) and Fisher and Gilles (1998) offer a continuous-time analogue. We adopt the Duffie and Epstein (1992b) parameterization:

$$u(C_t) = \frac{\beta}{1-1/\psi}(1-\gamma)J_t\left[\left(\frac{C_t}{((1-\gamma)J_t)^{1/(1-\gamma)}}\right)^{1-1/\psi} - 1\right], \tag{8}$$

where J_t is the indirect utility function defined by

$$J_t = E_t\left[\int_t^T u(C_v)\,\mathrm{d}v\right],$$

where $u(C)$ is therefore a normalized aggregator of both current consumption and continuation utility. $\beta > 0$ is the rate of time preference, $\gamma > 0$ is the coefficient of relative risk aversion and $\psi > 0$ is the elasticity of intertemporal substitution. Power utility obtains from (8) by setting $\psi = 1/\gamma$.

We also specify the utility of bequest, $Q(W_T)$, as

$$Q(W_T) = \frac{W_T^{1-\gamma}}{1-\gamma}.$$

[7] Power utility restricts the elasticity of intertemporal substitution parameter to be the inverse of the relative risk aversion coefficient, while in fact these two parameters need not be related to one another.

3. Exact solution for unit elasticity of intertemporal substitution

An exact solution to the model setup of the previous section is possible only in a handful of special cases. The most well known case is when the elasticity of substitution and the risk aversion coefficient both take the value of one. In this case, the investor's preferences are captured by a log utility function, where the time variation of the investment opportunity set is ignored in the consumption/investment decision.[8]

We show in this section that the reason log utility delivers an exact solution is primarily because log utility sets the value of the elasticity of substitution to one, not because it sets the risk aversion parameter to one. In other words, the difficulty involved with generating exact solutions lies in the elasticity of substitution parameter, not the risk aversion. Fixing the elasticity of substitution to one allows us to generate an exact solution to the consumption/investment problem for any risk aversion parameter. This is an extremely useful result for studying how risk aversion affects consumption/investment decisions as it greatly simplifies the analysis required.

3.1. General results

The normalized aggregator $u(C_t)$ takes the following form when $\psi \to 1$

$$u(C_t) = \beta(1-\gamma)J_t\left[\log(C_t) - \frac{1}{1-\gamma}\log\big((1-\gamma)J_t\big)\right]. \tag{9}$$

With this specification for the aggregator, the first-order conditions for (6) are

$$C_t = \frac{\beta(1-\gamma)J}{J_W}, \tag{10}$$

$$\pi_t = -\left(\frac{J_W}{WJ_{WW}}\right)\left(\frac{m-r}{v}\right) - \sum_{i=m,r,y}\left(\frac{J_{iW}}{WJ_{WW}}\right)\left(\frac{\rho_{iS}\sigma_i}{\sqrt{v}}\right). \tag{11}$$

Equation (10) results from the envelope condition, $\frac{\partial u}{\partial C} = \frac{\partial J}{\partial W}$, from which the optimal consumption rule obtains once the value function is known. Equation (11) shows that the optimal portfolio share in the risky asset has two components. The first one is the optimal demand for risky assets we find in single-period models, or in multi-period models with constant investment opportunities. For this reason it is called "myopic demand". The second component is Merton's intertemporal hedging demands. This component represents additional holdings in the risky asset in order to hedge against movements in the investment opportunity set. As such, this component depends critically on the

[8] Other cases where an exact solution is available is when there is no intermediate consumption (Liu, 1999), when the markets are complete, i.e., the investment opportunity set has a perfect positive or negative correlation with the risky asset (Wachter, 1999), or when changes in the investment opportunity set are completely unhedgeable, i.e., the opportunity set has zero correlation with the risky asset.

correlation between the risky asset and the state variables in the economy. By inspection of (11), it is immediately seen that the hedging component of portfolio demand is non-zero unless $\sigma = 0$ (constant investment opportunities), $\rho = 0$ (no hedging value in risky asset), or $J_{iS} = 0$, $i = m, r, y$. This last equality obtains when $\gamma = 1$ (Merton, 1969, 1971, 1973; Giovannini and Weil, 1989). Equations (10) and (11) do not represent a complete solution to the model until we solve for J.

Substituting the first-order conditions into (6) for the control variables and rearranging gives the Bellman equation

$$\begin{aligned} 0 = {} & \beta(1-\gamma)J\left[\log\left(\frac{\beta(1-\gamma)J}{J_W}\right) - \frac{\log(1-\gamma)}{1-\gamma} - \frac{\log J}{1-\gamma} - 1\right] \\ & + rWJ_W - \frac{(m-r)^2 J_W^2}{2vJ_{WW}} - \sum_{i=m,r,v} \frac{\rho_{iS}\sigma_i(m-r)J_{iW}J_W}{\sqrt{v}J_{WW}} \\ & - \sum_{i=m,r,y}\sum_{j=m,r,y} \frac{\rho_{iS}\rho_{jS}\sigma_i\sigma_j J_{iW}J_{jW}}{2J_{WW}} + \sum_{i=m,r,y}\mu_i J_i + \sum_{i=m,r,y}\frac{1}{2}\sigma_i^2 J_{ii} + J_t. \end{aligned} \tag{12}$$

The boundary condition for this nonlinear partial differential equation is given by $J(t = T) = \frac{W^{1-\gamma}}{1-\gamma}$, which states that the indirect utility at the end of the consumption/investment horizon must equal the utility from bequests. So at the end of the planning horizon, the agent bequests all remaining wealth and receives utility from doing so.

We now guess a solution for the Bellman equation of the form $J = I(r, v, m)W^{1-\gamma}/(1-\gamma)$ where $I(r, v, m)$ is a to-be-determined function of the state variables. Substituting this solution into the Bellman equation and simplifying yields the following partial differential equation (PDE)

$$\begin{aligned} 0 = {} & \beta\log\beta I - \frac{\beta}{1-\gamma} I\log I - \beta I + rI + \frac{(m-r)^2}{2\gamma v} I \\ & + \sum_{i=m,r,y}\frac{\rho_{iS}\sigma_i(m-r)}{\gamma\sqrt{v}} I_i + \sum_{i=m,r,y}\sum_{j=m,r,y}\frac{\rho_{iS}\rho_{jS}\sigma_i\sigma_j}{2\gamma}\frac{I_i I_j}{I} \\ & + \frac{1}{1-\gamma}\sum_{i=m,r,y}\mu i I_i + \frac{1}{1-\gamma}\sum_{i=m,r,y}\frac{1}{2}\sigma_i^2 I_{ii} + \frac{1}{1-\gamma} I_t. \end{aligned} \tag{13}$$

The subscripts on the function I represent partial derivatives. This equation is again a nonlinear partial differential equation, and the boundary condition is $I(t = T) = 1$.

For a general system of state variables, there is no known solution for this equation nor can it be proven that one exists. However, we can derive an exact, analytic solution when (1) the time-variation of volatility and expected returns can be captured entirely by the dynamics of the Sharpe ratio, and (2) the Sharpe ratio, the interest rate, and log stock returns form an affine system. The form of the solution is shown in the following proposition.

Proposition 2. *Assume that the time variation of volatility and expected returns can be captured entirely by the dynamics of the Sharpe ratio, $X = (m - r)/\sqrt{v}$. Then, if the dynamics of the Sharpe ratio*

$$\mathrm{d}X_t = \mu_X(X_t)\,\mathrm{d}t + \sigma_X(X_t)\,\mathrm{d}Z_{X,t}$$

and the dynamics of the interest rate

$$\mathrm{d}r_t = \mu_r(r_t)\,\mathrm{d}t + \sigma_r(r_t)\,\mathrm{d}Z_{r,t}$$

combined with the log stock return process form an affine system where the volatility of the Sharpe ratio is constant, then the solution to (13) must have the exponential affine form

$$I = \exp\bigl[A(\tau)X + B(\tau)X^2 + C(\tau)r + D(\tau)\bigr]$$

where $\tau = T - t$ represents the time difference between time t and the end of the investment horizon.

Proof. If the Sharpe ratio's dynamics capture the dynamics of the volatility and expected return of the risky asset, then (13) can be written as

$$\begin{aligned}0 = {}& \beta\log\beta I - \frac{\beta}{1-\gamma} I\log I - \beta I + rI + \frac{1}{2\gamma}X^2 I + \frac{\rho_{XS}\sigma_X}{\gamma} X I_X + \frac{\rho_{rS}\sigma_r}{\gamma} X I_r \\ &+ \frac{\rho_{XS}^2\sigma_X^2}{2\gamma}\frac{I_X^2}{I} + \frac{\rho_{XS}\rho_{rS}\sigma_X\sigma_r}{\gamma}\frac{I_X I_r}{I} + \frac{\rho_{rS}^2\sigma_r^2}{2\gamma}\frac{I_r^2}{I} + \frac{1}{1-\gamma}\mu_X I_X \\ &+ \frac{1}{1-\gamma}\mu_r I_r + \frac{1}{2(1-\gamma)}\sigma_X^2 I_{XX} + \frac{1}{2(1-\gamma)}\sigma_r^2 I_{rr} - \frac{1}{1-\gamma} I_\tau. \end{aligned} \tag{14}$$

If log stock returns, the Sharpe ratio, and the interest rate form a trivariate affine system, then σ_X and σ_r must be constants. Furthermore, μ_X and μ_r must be linear functions of X and r, respectively. Let

$$\mu_X = a_x + b_x X, \qquad \mu_r = a_r + b_r r.$$

Then, substituting the solution $I = \exp[A(\tau)X + B(\tau)X^2 + C(\tau)r + D(\tau)]$ into the PDE and rearranging, yields

$$\begin{aligned}0 = {}& \beta\log\beta - \frac{\beta}{1-\gamma}\bigl[A(\tau)X + B(\tau)X^2 + C(\tau)r + D(\tau)\bigr] - \beta + r + \frac{1}{2\gamma}X^2 \\ &+ \frac{\rho_{XS}\sigma_X}{\gamma}\bigl[A(\tau) + 2B(\tau)X\bigr]X + \frac{\rho_{rS}\sigma_r}{\gamma}C(\tau)X + \frac{\rho_{XS}^2\sigma_X^2}{2\gamma}\bigl[A(\tau) + 2B(\tau)X\bigr]^2 \\ &+ \frac{\rho_{XS}\rho_{rS}\sigma_X\sigma_r}{\gamma}\bigl[A(\tau) + 2B(\tau)X\bigr]C(\tau) + \frac{\rho_{rS}^2\sigma_r^2}{2\gamma}C(\tau)^2 \\ &+ \frac{1}{1-\gamma}[a_x + b_x X]\bigl[A(\tau) + 2B(\tau)X\bigr] + \frac{1}{1-\gamma}[a_r + b_r r]C(\tau)\end{aligned}$$

$$+\frac{1}{2(1-\gamma)}\sigma_X^2\big[A(\tau)+2B(\tau)X\big]^2+\frac{1}{(1-\gamma)}\sigma_X^2B(\tau)+\frac{1}{2(1-\gamma)}\sigma_r^2C(\tau)^2$$
$$-\frac{1}{1-\gamma}\left[\frac{\mathrm{d}A(\tau)}{\mathrm{d}\tau}X+\frac{\mathrm{d}B(\tau)}{\mathrm{d}\tau}X^2+\frac{\mathrm{d}C(\tau)}{\mathrm{d}\tau}r+\frac{\mathrm{d}D(\tau)}{\mathrm{d}\tau}\right]$$

which is a separable equation. Separating by variables gives the following system of four ordinary differential equations

$$\begin{aligned}
\frac{\mathrm{d}A(\tau)}{\mathrm{d}\tau} &= \left[\frac{(1-\gamma)\rho_{XS}\sigma_X}{\gamma}+\frac{2(1-\gamma)\rho_{XS}^2\sigma_X^2}{\gamma}B(\tau)+\sigma_X^2B(\tau)+b_x-\beta\right]A(\tau)\\
&\quad+2a_xB(\tau)+\frac{2(1-\gamma)\rho_{XS}\rho_{rS}\sigma_X\sigma_r}{\gamma}B(\tau)C(\tau)+\frac{(1-\gamma)\rho_{rS}\sigma_r}{\gamma}C(\tau),\\
\frac{\mathrm{d}B(\tau)}{\mathrm{d}\tau} &= \left[2\sigma_X^2+\frac{2(1-\gamma)\rho_{XS}^2\sigma_X^2}{\gamma}\right]B(\tau)^2\\
&\quad+\left[\frac{2(1-\gamma)\rho_{XS}\sigma_X}{\gamma}-\beta+2b_x\right]B(\tau)+\frac{(1-\gamma)}{2\gamma},\\
\frac{\mathrm{d}C(\tau)}{\mathrm{d}\tau} &= (b_r-\beta)C(\tau)+1-\gamma,\\
\frac{\mathrm{d}D(\tau)}{\mathrm{d}\tau} &= -\beta D(\tau)+\left[\frac{(1-\gamma)\rho_{XS}^2\sigma_X^2}{2\gamma}+\frac{1}{2}\sigma_X^2\right]A(\tau)^2\\
&\quad+\frac{(1-\gamma)\rho_{XS}\rho_{rS}\sigma_X\sigma_r}{\gamma}A(\tau)C(\tau)\\
&\quad+\left[\frac{(1-\gamma)\rho_{rS}^2\sigma_r^2}{2\gamma}+\frac{1}{2}\sigma_r^2\right]C(\tau)^2+a_xA(\tau)+\sigma_X^2B(\tau)\\
&\quad+a_rC(\tau)+\beta(1-\gamma)(\log\beta-1).
\end{aligned}$$

The equations for $A(\tau)$, $C(\tau)$, and $D(\tau)$ are first-order linear differential equations, while the equation for $B(\tau)$ is a Riccati equation. The boundary conditions for each equation are

$$A(0)=0,\qquad B(0)=0,\qquad C(0)=0,\qquad D(0)=0.$$

It is easy to show that solutions exist for each of these equations. To solve this system of equations, one would solve the Riccati equation for $B(\tau)$ and the simple ODE for $C(\tau)$ first. With these two solutions in hand, one would next solve the ODE for $A(\tau)$, as this equation requires the solutions for $B(\tau)$ and $C(\tau)$. Finally, $D(\tau)$ may be solved once the solutions for $A(\tau)$, $B(\tau)$, and $C(\tau)$ are obtained. □

From this proposition, we have the result that the indirect utility function has the form

$$J=\frac{1}{1-\gamma}\exp\big[A(\tau)X+B(\tau)X^2+C(\tau)r+D(\tau)\big]W^{1-\gamma}. \tag{15}$$

Consequently, the solutions for optimal consumption and portfolio choice are

$$C_t = \beta W_t, \tag{16}$$

$$\pi_t = \frac{1}{\gamma} X_t + \big[A(\tau) + 2B(\tau)X_t\big]\frac{\rho_{XS}\sigma_X}{\gamma\sqrt{v_t}} + C(\tau)\frac{\rho_{rS}\sigma_r}{\gamma\sqrt{v_t}}. \tag{17}$$

It is interesting that the function $D(\tau)$ does not need to be known in order to evaluate the optimal consumption and portfolio choice policies.

We now offer the following lemma for the infinite horizon setting.

Lemma 3. *If the conditions for Proposition 2 are satisfied for an infinite horizon consumption/investment problem, then the indirect utility function has the same form given in Proposition 2 except that it is time homogeneous:*

$$J_t = I(r, v, m)W^{1-\gamma}/(1-\gamma)$$

with

$$I(r, v, m) = \exp\big[AX + BX^2 + Cr + D\big]$$

where A, B, C, and D are constants.

Proof. The proof follows exactly along the lines of the proof for Proposition 2. □

3.2. Example 1: Time-varying expected returns (finite horizon)

We now illustrate the results of the previous section by solving the finite-horizon consumption/investment problem for the following specific form of the investment opportunity set

$$\frac{\mathrm{d}S_t}{S_t} = m_t\,\mathrm{d}t + \sigma\,\mathrm{d}Z_{S,t},$$
$$\mathrm{d}m_t = \kappa(\theta - m_t)\,\mathrm{d}t + \delta\,\mathrm{d}Z_{m,t}$$

where the correlation between stock returns and expected returns is a constant ρ_{mS}. The interest rate, r, is also assumed to be constant. The utility function is the recursive utility formulation with the elasticity of substitution set to one. This satisfies the conditions of Proposition 2 because the expected return process is affine, and since stock return volatility and the interest rate are constant, there is a one-to-one relationship between the Sharpe ratio and expected returns. The Sharpe ratio is trivially a sufficient statistic to describe the dynamics of expected returns.

The first-order conditions for consumption and portfolio choice are

$$C_t = \frac{\beta(1-\gamma)J}{J_W},$$

$$\pi_t = -\left(\frac{J_W}{WJ_{WW}}\right)\left(\frac{m-r}{\sigma^2}\right) - \left(\frac{J_{mW}}{WJ_{WW}}\right)\left(\frac{\rho_{mS}\delta}{\sigma}\right).$$

The Bellman equation is

$$0 = \beta(1-\gamma)J\left[\log\left(\frac{\beta(1-\gamma)J}{J_W}\right) - \frac{\log(1-\gamma)}{1-\gamma} - \frac{\log J}{1-\gamma} - 1\right] + rWJ_W$$
$$- \frac{(m-r)^2 J_W^2}{2\sigma^2 J_{WW}} - \frac{\rho_{mS}\delta(m-r)J_{mW}J_W}{\sigma J_{WW}} - \frac{\rho_{mS}^2\delta^2 J_{mW}^2}{2J_{WW}}$$
$$+ \kappa(\theta - m)J_m + \frac{1}{2}\delta^2 J_{mm} + J_t.$$

Guessing a solution of the form $J = I(m,t)W^{1-\gamma}/1-\gamma$ results in the PDE

$$0 = \beta\log\beta I - \frac{\beta}{1-\gamma}I\log I - \beta I + rI + \frac{(m-r)^2}{2\gamma\sigma^2}I + \frac{\rho_{mS}\delta(m-r)}{\gamma\sigma}I_m$$
$$+ \frac{\rho_{mS}^2\sigma_m^2}{2\gamma}\frac{I_m^2}{I} + \frac{1}{1-\gamma}\kappa(\theta-m)I_m + \frac{1}{2(1-\gamma)}\delta^2 I_{mm} + \frac{1}{1-\gamma}I_t.$$

Because the conditions of this problem satisfy those given Proposition 2, we guess an exponential affine solution form for I: $I = \exp[A(\tau)m + B(\tau)m^2 + D(\tau)]$. As shown in the proof to the proposition, substituting this solution form into the PDE results in three ordinary differential equations

$$\frac{\mathrm{d}A}{\mathrm{d}\tau} = \left[\left(\frac{2(1-\gamma)\rho_{mS}^2\sigma_m^2}{\gamma} + 2\delta^2\right)B + \frac{(1-\gamma)\rho_{mS}\delta}{\gamma\sigma} - \kappa - \beta\right]A$$
$$+ \left[2\kappa\theta + \delta^2 - \frac{2(1-\gamma)\rho_{mS}\delta}{\gamma\sigma}r\right]B - \frac{1-\gamma}{\gamma\sigma^2}r,$$
$$\frac{\mathrm{d}B}{\mathrm{d}\tau} = \frac{2(1-\gamma)\rho_{mS}^2\sigma_m^2}{\gamma}B^2 + \left[\frac{2(1-\gamma)\rho_{mS}\delta}{\gamma\sigma} - \beta\right]B + \frac{1-\gamma}{2\gamma\sigma^2},$$
$$\frac{\mathrm{d}D}{\mathrm{d}\tau} = -\beta C + \left[\frac{(1-\gamma)\rho_{mS}^2\sigma_m^2}{2\gamma} + \frac{1}{2}\delta^2\right]A^2 + \left[\kappa\theta - \frac{(1-\gamma)\rho_{mS}\delta}{\gamma\sigma}r\right]A$$
$$+ \beta(1-\gamma)(\log\beta - 1) + r + \frac{(1-\gamma)}{2\gamma\sigma^2}r^2.$$

The solution to $B(\tau)$ is

$$B(\tau) = \left(\frac{u_1u_2}{a_1}\right)\frac{\mathrm{e}^{u_1\tau} - \mathrm{e}^{u_2\tau}}{u_1\mathrm{e}^{u_2\tau} - u_2\mathrm{e}^{u_1\tau}}$$

with

$$u_1, u_2 = \frac{1}{2}b_1 \pm \sqrt{b_1^2 - 4a_1c_1},$$
$$a_1 = \frac{2(1-\gamma)\rho_{mS}^2\sigma_m^2}{\gamma},$$

$$b_1 = \frac{2(1-\gamma)\rho_{mS}\delta}{\gamma\sigma} - \beta,$$
$$c_1 = \frac{1-\gamma}{2\gamma\sigma^2}.$$

The solution to $A(\tau)$ can be calculated by evaluating the following integral:

$$A(\tau) = \frac{a_2}{a_1}\mathrm{e}^{b_2\tau}\left|\frac{u_2-u_1}{u_2\mathrm{e}^{u_1 v}-u_1\mathrm{e}^{u_2 v}}\right|^{a_2/a_1} \times \int_0^\tau \mathrm{e}^{-b_2 v}\left|\frac{u_2\mathrm{e}^{u_1 v}-u_1\mathrm{e}^{u_2 v}}{u_2-u_1}\right|^{a_2/a_1}\big(c_2 B(v)+d_2\big)\,\mathrm{d}v$$

with

$$a_2 = \frac{2(1-\gamma)\rho_{mS}^2\sigma_m^2}{\gamma} + 2\delta^2,$$
$$b_2 = \frac{(1-\gamma)\rho_{mS}\delta}{\gamma\sigma} - \kappa - \beta,$$
$$c_2 = 2\kappa\theta + \delta^2 - \frac{2(1-\gamma)\rho_{mS}\delta}{\gamma\sigma}r,$$
$$d_2 = -\frac{1-\gamma}{\gamma\sigma^2}r.$$

Finally, the solution to $D(\tau)$ can be calculated by evaluating another integral:

$$D(\tau) = \mathrm{e}^{\beta\tau}\int_0^\tau \mathrm{e}^{\beta v}\big[a_3 A(v)^2 + b_3 A(v)\big]\,\mathrm{d}v + \frac{c_3}{\beta}\big[\mathrm{e}^{2\beta\tau} - \mathrm{e}^{\beta\tau}\big]$$

with

$$a_3 = \frac{(1-\gamma)\rho_{mS}^2\sigma_m^2}{2\gamma} + \frac{1}{2}\delta^2,$$
$$b_3 = \kappa\theta - \frac{(1-\gamma)\rho_{mS}\delta}{\gamma\sigma}r,$$
$$c_3 = \beta(1-\gamma)(\log\beta - 1) + r + \frac{(1-\gamma)}{2\gamma\sigma^2}r^2.$$

Therefore, the optimal consumption and investment policies are

$$C_t = \beta W_t,$$
$$\pi_t = \frac{1}{\gamma}\left(\frac{m_t - r}{\sigma^2}\right) + \frac{\rho_{mS}\delta}{\gamma\sigma}\big[A(\tau) + 2B(\tau)m_t\big].$$

3.3. Example 2: Time-varying expected returns (infinite horizon)

We now consider the example of the previous section but set in an infinite horizon economy. In a finite-horizon setting, even though the functions $B(\tau)$ and $D(\tau)$ can be expressed analytically as a function of an integral, it is difficult to evaluate these functions in closed-form.[9] This problem is resolved in an infinite-horizon setting because the optimization problem becomes time homogeneous. As a result, the functions $A(\tau)$, $B(\tau)$, and $D(\tau)$ become constants rather than time-varying functions.

In the infinite horizon setting, the agent's problem is

$$\begin{aligned}
&\sup_{\pi,C} E_0\left[\int_0^\infty u(C_t)\,\mathrm{d}t\right]\\
&\text{s.t.}\quad \mathrm{d}W_t = \left[\pi_t m_t W_t + (1-\pi_t) r W_t - C_t\right]\mathrm{d}t + \pi_t W_t \sigma\,\mathrm{d}Z_{S,t},\\
&\qquad \frac{\mathrm{d}S_t}{S_t} = m_t\,\mathrm{d}t + \sigma\,\mathrm{d}Z_{S,t},\\
&\qquad \frac{\mathrm{d}B_t}{B_t} = r\,\mathrm{d}t
\end{aligned}$$

where the dynamics of the expected return are[10]

$$\mathrm{d}m_t = \kappa(\theta - m_t)\,\mathrm{d}t + \delta\,\mathrm{d}Z_{m,t}.$$

The expected utility from bequeathing wealth must be zero. Therefore, the boundary condition is the transversality condition

$$\lim_{t\to\infty} E_t(J_t) = 0.$$

The first-order condition is exactly the same as that derived in the finite horizon case, while the Bellman equation is also the same, with the exception of one term. Therefore, the indirect utility function has the form $J = I(m)W^{1-\gamma}/1-\gamma$ where $I(m)$ satisfies the ordinary differential equation

$$\begin{aligned}
0 = {}& \beta\log\beta I - \frac{\beta}{1-\gamma} I\log I - \beta I + rI + \frac{(m-r)^2}{2\gamma\sigma^2} I + \frac{\rho_{mS}\delta(m-r)}{\gamma\sigma} I_m\\
&+ \frac{\rho_{mS}^2\delta^2}{2\gamma}\frac{I_m^2}{I} + \frac{1}{1-\gamma}\kappa(\theta-m)I_m + \frac{1}{2(1-\gamma)}\delta^2 I_{mm}.
\end{aligned}$$

Because of the time homogeneity of the infinite horizon setting, the difference between this ODE and the corresponding PDE from the finite horizon case is that this equation contains no time derivative.

[9] To evaluate optimal consumption and investment decisions, the function $C(\tau)$ does not need to be evaluated.

[10] See Campbell et al. (2004) for more details on this problem.

From Lemma 3, the solution to this equation has the form

$$I = \exp\bigl[Am + Bm^2 + D\bigr]$$

where A, B, and D are constants. Substituting this solution form into the equations yields the following system of equations which can be solved trivially to find the values of A, B, and D.

$$\begin{aligned}
0 &= \left[\frac{2\rho_{mS}^2\sigma_m^2}{\gamma} + \frac{2\delta^2}{1-\gamma}\right]B^2 + \left[\frac{2\rho_{mS}\delta}{\gamma\sigma} - \frac{\beta}{1-\gamma} - \frac{2\kappa}{1-\gamma}\right]B + \frac{1}{2\gamma\sigma^2},\\
0 &= \left[\frac{\rho_{mS}^2\sigma_m^2}{\gamma}B + \frac{2\delta^2}{1-\gamma}B + \frac{\rho_{mS}\delta}{\gamma\sigma} - \frac{\beta}{1-\gamma} - \frac{\kappa}{1-\gamma}\right]A\\
&\quad + \left[\frac{2\kappa\theta}{1-\gamma} - \frac{2\rho_{mS}\delta r}{\gamma\sigma}\right]B - \frac{r}{\gamma\sigma^2},\\
0 &= -\frac{\beta}{1-\gamma}D + \left[\frac{\rho_{mS}^2\sigma_m^2}{2\gamma} + \frac{\delta^2}{2(1-\gamma)}\right]A^2 - \left[\frac{\kappa\theta}{1-\gamma} + \frac{\rho_{mS}\delta r}{\gamma\sigma}\right]A + \frac{\delta^2}{1-\gamma}B\\
&\quad + \beta(\log\beta - 1) + r + \frac{r^2}{2\gamma\sigma^2}.
\end{aligned}$$

From the first-order conditions, the optimal consumption and investment policies are

$$\begin{aligned}
C_t &= \beta W_t,\\
\pi_t &= \frac{1}{\gamma}\left(\frac{m_t - r}{\sigma^2}\right) + \frac{\rho_{mS}\delta}{\gamma\sigma}[A + 2Bm_t].
\end{aligned}$$

4. Approximate solution for general elasticity of intertemporal substitution

Thus far, we have solved the consumption/investment problem for a recursive utility agent with a unit elasticity of substitution. This section deals with the case when the general case, when the elasticity of substitution is not one. In this general case, an exact solution is infeasible. Therefore, we offer two approaches to deriving approximate solutions for optimal consumption and portfolio choice. The first is a perturbation to the solution obtained with unit elasticity of substitution. This solution works best when the elasticity of substitution is close to one. The second method, which works well even when the elasticity of substitution is far from one, is to take a perturbation around the unconditional mean of the optimal consumption/wealth ratio.

4.1. Perturbation around unit elasticity of substitution

In the previous section, it was demonstrated in Proposition 2 and Lemma 3 that an exact, analytic solution existed for a class of consumption/investment problems when the

elasticity of substitution takes the value of one. We build upon these results in this section by deriving an approximation technique that involves taking an expansion around the exact solution generated for the unit elasticity of substitution case. Because an exact solution for unit elasticity of substitution is available in both the finite horizon and infinite horizon cases, an approximate solution is available in both these cases as well. As with the exact solutions, the approximate solutions are also separable in wealth and other state variables, where the state variable function is exponential-affine.

We begin by restating the general consumption/investment problem:

$$\begin{aligned}
&\sup_{\pi,C} E_0\left[\int_0^T \frac{\beta(1-\gamma)J_t}{1-1/\psi}\left[\left(\frac{C_t}{((1-\gamma)J_t)^{1/(1-\gamma)}}\right)^{1-1/\psi}-1\right]\mathrm{d}t+\frac{W_T^{1-\gamma}}{1-\gamma}\right],\\
&\text{s.t.}\quad \mathrm{d}W_t=\left[\pi_t m_t W_t+(1-\pi_t)r_t W_t-C_t\right]\mathrm{d}t+\pi_t W_t\sqrt{v_t}\,\mathrm{d}Z_{S,t},\\
&\qquad \frac{\mathrm{d}S_t}{S_t}=m_t\,\mathrm{d}t+\sqrt{v_t}\,\mathrm{d}Z_{S,t},\\
&\qquad \frac{\mathrm{d}B_t}{B_t}=r_t\,\mathrm{d}t.
\end{aligned}\tag{18}$$

The dynamics of the investment opportunity set are given by (5). The Sharpe ratio's dynamics are specified by (7).

The Principle of Optimality leads to the same equation derived in (6). From here the first-order conditions are

$$\begin{aligned}
C&=(J_W)^{-\psi}\left[(1-\gamma)J\right]^{(1-\gamma\psi)/(1-\gamma)}\beta^{\psi},\\
\pi_t&=-\left(\frac{J_W}{WJ_{WW}}\right)\left(\frac{m-r}{v}\right)-\sum_{i=m,r,y}\left(\frac{J_{iW}}{WJ_{WW}}\right)\left(\frac{\rho_{iS}\sigma_i}{\sqrt{v}}\right).
\end{aligned}$$

While the first-order condition for optimal portfolio choice remains the same as in the unit elasticity of substitution case, the first-order condition for optimal consumption becomes more complicated. It is the additional complications of this first-order condition that causes difficulty in solving consumption/investment problems.

These first-order conditions lead to the Bellman equation:

$$\begin{aligned}
0=&\frac{\beta(1-\gamma)}{1-1/\psi}J\left[\left(\frac{(J_W)^{-\psi}[(1-\gamma)J]^{(1-\gamma\psi)/(1-\gamma)}\beta^{\psi}}{((1-\gamma)J)^{1/(1-\gamma)}}\right)^{1-1/\psi}-1\right]\\
&-(J_W)^{1-\psi}\left[(1-\gamma)J\right]^{(1-\gamma\psi)/(1-\gamma)}\beta^{\psi}+rWJ_W-\frac{(m-r)^2J_W^2}{2vJ_{WW}}\\
&-\sum_{i=m,r,v}\frac{\rho_{iS}\sigma_i(m-r)J_{iW}J_W}{\sqrt{v}J_{WW}}-\sum_{i=m,r,y}\sum_{j=m,r,y}\frac{\rho_{iS}\rho_{jS}\sigma_i\sigma_jJ_{iW}J_{jW}}{2J_{WW}}\\
&+\sum_{i=m,r,y}\mu_iJ_i+\sum_{i=m,r,y}\frac{1}{2}\sigma_i^2J_{ii}+J_t.
\end{aligned}\tag{19}$$

The boundary condition for this nonlinear PDE is given by $J(t=T)=\frac{W^{1-\gamma}}{1-\gamma}$.

We now guess a solution for the Bellman equation of the form $J = I(r, v, m)W^{1-\gamma}/(1-\gamma)$ where, as in the previous section, $I(r, v, m)$ is a to-be-determined function of the state variables. Substituting this solution into the Bellman equation and simplifying yields the following partial differential equation (PDE):

$$\begin{aligned} 0 = {} & \frac{\beta^{\psi}}{\psi - 1} I^{(1-\psi)/(1-\gamma)+1} - \frac{\psi\beta}{\psi - 1} I + rI + \frac{(m-r)^2}{2\gamma v} I \\ & + \sum_{i=m,r,y} \frac{\rho_{iS}\sigma_i(m-r)}{\gamma\sqrt{v}} I_i + \sum_{i=m,r,y} \sum_{j=m,r,y} \frac{\rho_{iS}\rho_{jS}\sigma_i\sigma_j}{2\gamma} \frac{I_i I_j}{I} \\ & + \frac{1}{1-\gamma} \sum_{i=m,r,y} \mu_i I_i + \frac{1}{1-\gamma} \sum_{i=m,r,y} \frac{1}{2}\sigma_i^2 I_{ii} + \frac{1}{1-\gamma} I_t \end{aligned} \tag{20}$$

with the boundary condition $I(t = T) = 1$. Analogous to the case of unit elasticity of substitution, we now have the following proposition which shows that an approximate, exponential-affine solution is available for I when (1) the time-variation of volatility and expected returns can be captured entirely by the dynamics of the Sharpe ratio, and (2) the Sharpe ratio, the interest rate, and log stock returns form an affine system.

Proposition 4. *Assume that the time variation of volatility and expected returns can be captured entirely by the dynamics of the Sharpe ratio, $X = (m - r)/\sqrt{v}$. Then, if the dynamics of the Sharpe ratio*

$$\mathrm{d}X_t = \mu_X(X_t)\,\mathrm{d}t + \sigma_X(X_t)\,\mathrm{d}Z_{X,t}$$

and the dynamics of the interest rate

$$\mathrm{d}r_t = \mu_r(r_t)\,\mathrm{d}t + \sigma_r(r_t)\,\mathrm{d}Z_{r,t}$$

combined with the log stock return process form an affine system where the volatility of the Sharpe ratio is constant, then an approximate solution to (20) has the exponential affine form given by

$$I = \exp\bigl[A(\tau)X + B(\tau)X^2 + C(\tau)r + D(\tau)\bigr]$$

where $\tau = T - t$ represents the time difference between time t and the end of the investment horizon.

Proof. If the Sharpe ratio's dynamics capture the dynamics of the volatility and expected return of the risky asset, then (20) can be written as

$$\begin{aligned} 0 = {} & \frac{\beta^{\psi}}{\psi - 1} I^{(1-\psi)/(1-\gamma)+1} - \frac{\psi\beta}{\psi - 1} I + rI + \frac{1}{2\gamma} X^2 I + \frac{\rho_{XS}\sigma_X}{\gamma} X I_X \\ & + \frac{\rho_{rS}\sigma_r}{\gamma} X I_r + \frac{\rho_{XS}^2\sigma_X^2}{2\gamma} \frac{I_X^2}{I} + \frac{\rho_{XS}\rho_{rS}\sigma_X\sigma_r}{\gamma} \frac{I_X I_r}{I} + \frac{\rho_{rS}^2\sigma_r^2}{2\gamma} \frac{I_r^2}{I} \end{aligned}$$

$$+\frac{1}{1-\gamma}\mu_X I_X+\frac{1}{1-\gamma}\mu_r I_r+\frac{1}{2(1-\gamma)}\sigma_X^2 I_{XX}$$
$$+\frac{1}{2(1-\gamma)}\sigma_r^2 I_{rr}-\frac{1}{1-\gamma}I_\tau.$$

One term, the first term on the right-hand side of the above equation, creates all of the difficulty in solving the above equation. Consequently, we take a first-order series expansion of this one term:

$$I^{1+(1-\psi)/(1-\gamma)}\approx I+\frac{1-\psi}{1-\gamma}I\log I.$$

The resulting equation is

$$0=\frac{\beta^\psi-\psi\beta}{\psi-1}I-\frac{\beta^\psi}{1-\gamma}I\log I+rI+\frac{1}{2\gamma}X^2 I+\frac{\rho_{XS}\sigma_X}{\gamma}XI_X+\frac{\rho_{rS}\sigma_r}{\gamma}XI_r$$
$$+\frac{\rho_{XS}^2\sigma_X^2}{2\gamma}\frac{I_X^2}{I}+\frac{\rho_{XS}\rho_{rS}\sigma_X\sigma_r}{\gamma}\frac{I_X I_r}{I}+\frac{\rho_{rS}^2\sigma_r^2}{2\gamma}\frac{I_r^2}{I}+\frac{1}{1-\gamma}\mu_X I_X$$
$$+\frac{1}{1-\gamma}\mu_r I_r+\frac{1}{2(1-\gamma)}\sigma_X^2 I_{XX}+\frac{1}{2(1-\gamma)}\sigma_r^2 I_{rr}-\frac{1}{1-\gamma}I_\tau.$$

It is interesting that if we take the limit as $\psi\to 1$ the above PDE becomes exactly the same as the corresponding PDE, (14), for the case with elasticity of substitution equal to one.

Since the PDE now is only differs from (14) by two constant terms, we can solve this PDE analogously to the solution used in the proof to Proposition 2. If we substitute the solution form $I=\exp[A(\tau)X+B(\tau)X^2+C(\tau)r+D(\tau)]$ into the PDE and separate variables, we get the following system of ODEs for the functions $A(\tau)$, $B(\tau)$, $C(\tau)$, and $D(\tau)$

$$\frac{\mathrm{d}A(\tau)}{\mathrm{d}\tau}=\left[\frac{(1-\gamma)\rho_{XS}\sigma_X}{\gamma}+\frac{2(1-\gamma)\rho_{XS}^2\sigma_X^2}{\gamma}B(\tau)+\sigma_X^2 B(\tau)+b_x-\beta^\psi\right]A(\tau)$$
$$+2a_x B(\tau)+\frac{2(1-\gamma)\rho_{XS}\rho_{rS}\sigma_X\sigma_r}{\gamma}B(\tau)C(\tau)+\frac{(1-\gamma)\rho_{rS}\sigma_r}{\gamma}C(\tau),$$
$$\frac{\mathrm{d}B(\tau)}{\mathrm{d}\tau}=\left[2\sigma_X^2+\frac{2(1-\gamma)\rho_{XS}^2\sigma_X^2}{\gamma}\right]B(\tau)^2$$
$$+\left[\frac{2(1-\gamma)\rho_{XS}\sigma_X}{\gamma}-\beta^\psi+2b_x\right]B(\tau)+\frac{(1-\gamma)}{2\gamma},$$
$$\frac{\mathrm{d}C(\tau)}{\mathrm{d}\tau}=\left(b_r-\beta^\psi\right)C(\tau)+1-\gamma,$$

$$\begin{aligned}\frac{\mathrm{d}D(\tau)}{\mathrm{d}\tau} &= -\beta^{\psi} D(\tau) + \left[\frac{(1-\gamma)\rho_{XS}^2\sigma_X^2}{2\gamma} + \frac{1}{2}\sigma_X^2\right] \\ &\quad \times A(\tau)^2 + \frac{(1-\gamma)\rho_{XS}\rho_{rS}\sigma_X\sigma_r}{\gamma} A(\tau)C(\tau) \\ &\quad + \left[\frac{(1-\gamma)\rho_{rS}^2\sigma_r^2}{2\gamma} + \frac{1}{2}\sigma_r^2\right] C(\tau)^2 + a_x A(\tau) + \sigma_X^2 B(\tau) \\ &\quad + a_r C(\tau) - \frac{1-\gamma}{1-\psi}\big(\beta^{\psi} - \psi\beta\big).\end{aligned}$$

The equations are very similar to those derived for the unit elasticity of substitution case, with the only changes being the replacement of a few constants with a different set of constants. In addition, the boundary condition for each equation are also the same as that for the unit elasticity of substitution case:

$$A(0) = 0, \qquad B(0) = 0, \qquad C(0) = 0, \qquad D(0) = 0.$$

As a result, not only do solutions exist for each of these equations, but the solutions are the same as those for the unit elasticity of substitution case but with a slightly different set of constants.

To solve this system of equations, one would solve the Riccati equation for $B(\tau)$ and the simple ODE for $C(\tau)$ first. With these two solutions in hand, one would next solve the ODE for $A(\tau)$, as this equation requires the solutions for $B(\tau)$ and $C(\tau)$. Finally, $D(\tau)$ may be solved once the solutions for $A(\tau)$, $B(\tau)$, and $C(\tau)$ are obtained. □

From this proposition, we see that the indirect utility function and optimal portfolio choice have the same form as those given by (15) and (17). Optimal consumption is

$$C_t = \beta^{\psi} \mathrm{e}^{[A(\tau)X_t + B(\tau)X_t^2 + C(\tau)r_t + D(\tau)]\frac{1-\psi}{1-\gamma}} W_t.$$

The following lemma gives the infinite horizon version of this proposition.

Lemma 5. *If the conditions for Proposition 4 are satisfied for an infinite horizon consumption/investment problem, then the indirect utility function has the same form given in Proposition 4 except that it is time homogeneous*

$$J_t = I(r, v, m)W^{1-\gamma}/(1-\gamma)$$

with

$$I(r, v, m) = \exp\big[AX + BX^2 + Cr + D\big]$$

where A, B, C, and D are constants.

Proof. The proof follows exactly along the lines of the proof for Proposition 4. □

We now rework the example of Section 3.2 for the case of arbitrary elasticity of substitution in order to illustrate the use of the results in this section. The stock dynamics are

$$\frac{\mathrm{d}S_t}{S_t} = m_t\,\mathrm{d}t + \sigma\,\mathrm{d}Z_{S,t},$$
$$\mathrm{d}m_t = \kappa(\theta - m_t)\,\mathrm{d}t + \delta\,\mathrm{d}Z_{m,t}.$$

The first-order conditions for the consumption/investment problem, (18), with recursive utility are

$$C = (J_W)^{-\psi}\big[(1-\gamma)J\big]^{(1-\gamma\psi)/(1-\gamma)}\beta^{\psi},$$
$$\pi_t = -\left(\frac{J_W}{WJ_{WW}}\right)\left(\frac{m-r}{\sigma^2}\right) - \left(\frac{J_{iW}}{WJ_{WW}}\right)\left(\frac{\rho_{mS}\delta}{\sigma}\right)$$

which leads to the Bellman equation:

$$\begin{aligned}0 = {} & \frac{\beta(1-\gamma)}{1-1/\psi}J\left[\left(\frac{(J_W)^{-\psi}[(1-\gamma)J]^{(1-\gamma\psi)/(1-\gamma)}\beta^{\psi}}{((1-\gamma)J)^{1/(1-\gamma)}}\right)^{1-1/\psi} - 1\right] \\ & - (J_W)^{1-\psi}\big[(1-\gamma)J\big]^{(1-\gamma\psi)/(1-\gamma)}\beta^{\psi} + rWJ_W \\ & - \frac{(m-r)^2J_W^2}{2\sigma^2 J_{WW}} - \frac{\rho_{mS}\delta(m-r)J_{mW}J_W}{\sigma J_{WW}} \\ & - \frac{\rho_{mS}^2\delta^2 J_{mW}^2}{2J_{WW}} + \kappa(\theta-m)J_m + \frac{1}{2}\delta^2 J_{mm} + J_t.\end{aligned}$$

Guessing the solution form $J = IW^{1-\gamma}/(1-\gamma)$ leads to the PDE

$$\begin{aligned}0 = {} & \frac{\beta^{\psi}}{\psi-1}I^{(1-\psi)/(1-\gamma)+1} - \frac{\psi\beta}{\psi-1}I + rI + \frac{(m-r)^2}{2\gamma\sigma^2}I + \frac{\rho_{mS}\delta(m-r)}{\gamma\sigma}I_m \\ & + \frac{\rho_{mS}^2\sigma_m^2}{2\gamma}\frac{I_m^2}{I} + \frac{1}{1-\gamma}\kappa(\theta-m)I_m + \frac{1}{2(1-\gamma)}\delta^2 I_{mm} + \frac{1}{1-\gamma}I_t.\end{aligned}$$

We now take a first-order series expansion of the term $I^{1+(1-\psi)/(1-\gamma)}$ around $\psi = 1$:

$$I^{1+(1-\psi)/(1-\gamma)} \approx I + \frac{1-\psi}{1-\gamma}I\log I.$$

Substituting this back into the PDE and rearranging gives

$$\begin{aligned}0 = {} & \frac{\beta^{\psi} - \psi\beta}{\psi-1}I - \frac{\beta^{\psi}}{1-\gamma}I\log I + rI + \frac{(m-r)^2}{2\gamma\sigma^2}I + \frac{\rho_{mS}\delta(m-r)}{\gamma\sigma}I_m \\ & + \frac{\rho_{mS}^2\sigma_m^2}{2\gamma}\frac{I_m^2}{I} + \frac{1}{1-\gamma}\kappa(\theta-m)I_m + \frac{1}{2(1-\gamma)}\delta^2 I_{mm} + \frac{1}{1-\gamma}I_t.\end{aligned}$$

Using the results of Proposition 4, we know that an approximate solution to the resulting equation has the form $I = \exp[A(\tau)m + B(\tau)m^2 + D(\tau)]$, which is the same as the unit

elasticity of substitution case. By analogy with the solution in Section 3.2, the solution to the indirect utility function can be written down by inspection

$$J_t = \exp\bigl[A(\tau)m_t + B(\tau)m_t^2 + D(\tau)\bigr]\frac{W_t^{1-\gamma}}{1-\gamma}$$

where

$$B(\tau) = \left(\frac{u_1u_2}{a_1}\right)\frac{\mathrm{e}^{u_1\tau} - \mathrm{e}^{u_2\tau}}{u_1\mathrm{e}^{u_2\tau} - u_2\mathrm{e}^{u_1\tau}},$$

$$A(\tau) = \frac{a_2}{a_1}\mathrm{e}^{b_2\tau}\left|\frac{u_2 - u_1}{u_2\mathrm{e}^{u_1 v} - u_1\mathrm{e}^{u_2 v}}\right|^{a_2/a_1} \times \int_0^\tau \mathrm{e}^{-b_2 v}\left|\frac{u_2\mathrm{e}^{u_1 v} - u_1\mathrm{e}^{u_2 v}}{u_2 - u_1}\right|^{a_2/a_1}\bigl(c_2B(v) + d_2\bigr)\,\mathrm{d}v,$$

$$D(\tau) = \mathrm{e}^{\beta^\psi\tau}\int_0^\tau \mathrm{e}^{\beta^\psi v}\bigl[a_3A(v)^2 + b_3A(v)\bigr]\,\mathrm{d}v + \frac{c_3}{\beta^\psi}\bigl[\mathrm{e}^{2\beta^\psi\tau} - \mathrm{e}^{\beta^\psi\tau}\bigr]$$

with

$$u_1, u_2 = \frac{1}{2}b_1 \pm \sqrt{b_1^2 - 4a_1c_1},$$

$$a_1 = \frac{2(1-\gamma)\rho_{mS}^2\sigma_m^2}{\gamma},$$

$$b_1 = \frac{2(1-\gamma)\rho_{mS}\delta}{\gamma\sigma} - \beta^\psi,$$

$$c_1 = \frac{1-\gamma}{2\gamma\sigma^2},$$

$$a_2 = \frac{2(1-\gamma)\rho_{mS}^2\sigma_m^2}{\gamma} + 2\delta^2,$$

$$b_2 = \frac{(1-\gamma)\rho_{mS}\delta}{\gamma\sigma} - \kappa - \beta^\psi,$$

$$c_2 = 2\kappa\theta + \delta^2 - \frac{2(1-\gamma)\rho_{mS}\delta}{\gamma\sigma}r,$$

$$d_2 = -\frac{1-\gamma}{\gamma\sigma^2}r,$$

$$a_3 = \frac{(1-\gamma)\rho_{mS}^2\sigma_m^2}{2\gamma} + \frac{1}{2}\delta^2,$$

$$b_3 = \kappa\theta - \frac{(1-\gamma)\rho_{mS}\delta}{\gamma\sigma}r,$$

$$c_3 = -\frac{1-\gamma}{1-\psi}\bigl(\beta^\psi - \psi\beta\bigr) + r + \frac{(1-\gamma)}{2\gamma\sigma^2}r^2.$$

Therefore, the optimal consumption and investment policies are

$$C_t = \beta^{\psi} \mathrm{e}^{[A(\tau)m_t + B(\tau)m_t^2 + D(\tau)]\frac{1-\psi}{1-\gamma}} W_t,$$
$$\pi_t = \frac{1}{\gamma}\left(\frac{m_t - r}{\sigma^2}\right) + \frac{\rho_{mS}\delta}{\gamma\sigma}\big[A(\tau) + 2B(\tau)m_t\big].$$

An immediate difference between this solution and that with unit elasticity of substitution is that consumption is no longer a constant proportion of wealth.

We can also quickly calculate the solution to the infinite horizon problem in Section 3.3 with arbitrary elasticity of substitution. If we take the limit of the finite horizon indirect utility function as the horizon goes to infinity, we get

$$J_t = \exp\big[Am_t + Bm_t^2 + D\big]\frac{W_t^{1-\gamma}}{1-\gamma}$$

where analogous to the solution in Section 3.3, A, B, and D solve the simple system of equations

$$0 = \left[\frac{2\rho_{mS}^2\sigma_m^2}{\gamma} + \frac{2\delta^2}{1-\gamma}\right]B^2 + \left[\frac{2\rho_{mS}\delta}{\gamma\sigma} - \frac{\beta^{\psi}}{1-\gamma} - \frac{2\kappa}{1-\gamma}\right]B + \frac{1}{2\gamma\sigma^2},$$
$$0 = \left[\frac{\rho_{mS}^2\sigma_m^2}{\gamma}B + \frac{2\delta^2}{1-\gamma}B + \frac{\rho_{mS}\delta}{\gamma\sigma} - \frac{\beta^{\psi}}{1-\gamma} - \frac{\kappa}{1-\gamma}\right]A + \left[\frac{2\kappa\theta}{1-\gamma} - \frac{2\rho_{mS}\delta r}{\gamma\sigma}\right]B - \frac{r}{\gamma\sigma^2},$$
$$0 = -\frac{\beta^{\psi}}{1-\gamma}D + \left[\frac{\rho_{mS}^2\sigma_m^2}{2\gamma} + \frac{\delta^2}{2(1-\gamma)}\right]A^2 + \left[\frac{\rho_{mS}\delta r}{\gamma\sigma} + \frac{\kappa\theta}{1-\gamma}\right]A + \frac{\delta^2}{1-\gamma}B - \frac{1-\gamma}{1-\psi}\big(\beta^{\psi} - \psi\beta\big) + r + \frac{r^2}{2\gamma\sigma^2}.$$

4.2. Perturbation around mean of consumption/wealth ratio

When the elasticity of substitution is close to one, a perturbation around the unit elasticity of substitution solution can give a fairly accurate solution. However, when the elasticity of substitution is far from one, this method cannot be relied upon. Consequently, we present a second technique for solving problems when the elasticity of substitution is not close to one. This technique involves taking an expansion around the unconditional mean of the consumption/wealth ratio.

The first order condition and Bellman equation are the same as those derived in the previous section. Therefore, if we guess a solution for the Bellman equation of the form $J = I(r, v, m)W^{1-\gamma}/(1-\gamma)$ where, as in the previous section, $I(r, v, m)$ is a to-be-determined function of the state variables. This results in the same equation as (20).

Analogous to the case of unit elasticity of substitution, we now have the following proposition which shows that an approximate, exponential-affine solution is available

for I when (1) the time-variation of volatility and expected returns can be captured entirely by the dynamics of the Sharpe ratio, and (2) the Sharpe ratio, the interest rate, and log stock returns form an affine system.

Proposition 6. *Assume that the time variation of volatility and expected returns can be captured entirely by the dynamics of the Sharpe ratio, $X = (m - r)/\sqrt{v}$. Then, if the dynamics of the Sharpe ratio*

$$\mathrm{d}X_t = \mu_X(X_t)\,\mathrm{d}t + \sigma_X(X_t)\,\mathrm{d}Z_{X,t}$$

and the dynamics of the interest rate

$$\mathrm{d}r_t = \mu_r(r_t)\,\mathrm{d}t + \sigma_r(r_t)\,\mathrm{d}Z_{r,t}$$

combined with the log stock return process form an affine system where the volatility of the Sharpe ratio is constant, then an approximate solution to (20) has the exponential affine form

$$I = \exp\left\{-\frac{1-\gamma}{1-\psi}\big[A(\tau)X + B(\tau)X^2 + C(\tau)r + D(\tau)\big]\right\}$$

where $\tau = T - t$ represents the time difference between time t and the end of the investment horizon.

Proof. If the Sharpe ratio's dynamics capture the dynamics of the volatility and expected return of the risky asset, then (20) can be written as

$$\begin{aligned}
0 = {} & \frac{\beta^{\psi}}{\psi-1} I^{(1-\psi)/(1-\gamma)+1} - \frac{\psi\beta}{\psi-1} I + rI + \frac{1}{2\gamma} X^2 I + \frac{\rho_{XS}\sigma_X}{\gamma} X I_X \\
& + \frac{\rho_{rS}\sigma_r}{\gamma} X I_r + \frac{\rho_{XS}^2\sigma_X^2}{2\gamma}\frac{I_X^2}{I} + \frac{\rho_{XS}\rho_{rS}\sigma_X\sigma_r}{\gamma}\frac{I_X I_r}{I} + \frac{\rho_{rS}^2\sigma_r^2}{2\gamma}\frac{I_r^2}{I} \\
& + \frac{1}{1-\gamma}\mu_X I_X + \frac{1}{1-\gamma}\mu_r I_r + \frac{1}{2(1-\gamma)}\sigma_X^2 I_{XX} \\
& + \frac{1}{2(1-\gamma)}\sigma_r^2 I_{rr} - \frac{1}{1-\gamma} I_{\tau}.
\end{aligned}$$

We now make the transformation

$$I = H^{-(1-\gamma)/(1-\psi)}.$$

The derivatives of I are

$$\begin{aligned}
I_X &= -\frac{1-\gamma}{1-\psi} H^{-(1-\gamma)/(1-\psi)-1} H_X, \\
I_r &= -\frac{1-\gamma}{1-\psi} H^{-(1-\gamma)/(1-\psi)-1} H_r,
\end{aligned}$$

$$I_{XX} = \frac{1-\gamma}{1-\psi}\left(\frac{1-\gamma}{1-\psi}+1\right)H^{-(1-\gamma)/(1-\psi)-2}(H_X)^2 - \frac{1-\gamma}{1-\psi}H^{-(1-\gamma)/(1-\psi)-1}H,$$

$$I_{rr} = \frac{1-\gamma}{1-\psi}\left(\frac{1-\gamma}{1-\psi}+1\right)H^{-(1-\gamma)/(1-\psi)-2}(H_r)^2 - \frac{1-\gamma}{1-\psi}H^{-(1-\gamma)/(1-\psi)-1}H.$$

The resulting equation is

$$\begin{aligned} 0 = {} & -\beta^{\psi}H^{-1} + \psi\beta + r(1-\psi) + \frac{1-\psi}{2\gamma}X^2 - \frac{\rho_{XS}\sigma_X(1-\gamma)}{\gamma}X\frac{H_X}{H} \\ & - \frac{\rho_{rS}\sigma_r(1-\gamma)}{\gamma}X\frac{H_r}{H} + \frac{\rho_{XS}^2\sigma_X^2(1-\gamma)^2}{2\gamma(1-\psi)}\frac{H_X^2}{H^2} \\ & + \frac{\rho_{XS}\rho_{rS}\sigma_X\sigma_r(1-\gamma)^2}{\gamma(1-\psi)}\frac{H_XH_r}{H^2} + \frac{\rho_{rS}^2\sigma_r^2(1-\gamma)^2}{2\gamma(1-\psi)}\frac{H_r^2}{H^2} - \mu_X\frac{H_X}{H} - \mu_r\frac{H_r}{H} \\ & + \frac{H_\tau}{H} + \frac{1}{2}\sigma_X^2\left[\left(\frac{1-\gamma}{1-\psi}+1\right)\frac{H_X^2}{H^2}-1\right] + \frac{1}{2}\sigma_r^2\left[\left(\frac{1-\gamma}{1-\psi}+1\right)\frac{H_r^2}{H^2}-1\right] \end{aligned}$$

with the boundary condition $H(\tau = 0) = 1$.

The above is difficult to solve due to the first term, $-\beta^{\psi}H^{-1}$. Therefore, the approach in this section is to linearize this one term around the unconditional mean of the consumption-to-wealth ratio. The envelope condition implies

$$\beta^{\psi}H_t^{-1} = \exp\{c_t - w_t\},$$

where $c_t - w_t = \log(C_t/W_t)$. Therefore, we can employ the following perturbation:

$$\begin{aligned} \beta^{\psi}H_t^{-1} &\approx \exp\{\overline{c-w}\} + \exp\{\overline{c-w}\}\big[(c_t - w_t) - (\overline{c-w})\big] \\ &\equiv h_0 + h_1(c_t - w_t), \end{aligned}$$

where $(\overline{c-w}) = \mathrm{E}[c_t - w_t]$, $h_0 = \exp\{\overline{c-w}\}[1 - (\overline{c-w})]$ and $h_1 = \exp\{\overline{c-w}\}$. Substituting the linear approximation for $\beta^{\psi}H_t^{-1}$ in the first term of the Bellman equation, it is easy to verify by using the same approach taken in proofs to the previous propositions that the resulting ODE has a solution of the form

$$H_t = \exp\big[A(\tau)m_t^2 + B(\tau)m_t + C(\tau)r_t + D(\tau)\big].$$

This solution implies that $c_t - w_t = -A(\tau)m_t^2 - B(\tau)m_t - C(\tau) - D(\tau) + \psi\log\beta$. □

The following lemma now gives the infinite horizon result.

Lemma 7. *If the conditions for Proposition 6 are satisfied for an infinite horizon consumption/investment problem, then the indirect utility function has the same form given*

in Proposition 6 *except that it is time homogeneous*

$$J_t \simeq H_t^{-(1-\gamma)/(1-\psi)} \frac{W^{1-\gamma}}{1-\gamma}$$

with

$$H = \exp\bigl[AX_t + BX_t^2 + Cr_t + D\bigr]$$

where A, B, C, and D are constants.

Proof. The proof follows exactly along the lines of the proof for Proposition 6 but with no time derivatives. □

We now solve the infinite horizon example done in Section 3.3 and the previous section to illustrate the technique.

Stock returns are assumed to have following properties

$$\frac{\mathrm{d}S_t}{S_t} = m_t\,\mathrm{d}t + \sigma\,\mathrm{d}Z_{S,t}, \qquad \mathrm{d}m_t = \kappa(\theta - m_t)\,\mathrm{d}t + \delta\,\mathrm{d}Z_{m,t}.$$

Thus, stock returns exhibit predictability in returns. The instantaneous correlation between $Z_{S,t}$ and $Z_{m,t}$ is given by ρ.

From the Principle of Optimality, the indirect utility function for the optimization problem satisfies the equation

$$\begin{aligned} 0 = \sup_{\pi,C}\Bigl\{ & \mathrm{f}(C, J) + \bigl[\pi(m-r)W + rW - C\bigr]J_W + \frac{1}{2}\pi^2 W^2\sigma^2 J_{WW} \\ & + \kappa(\theta - m)J_m + \frac{1}{2}\delta^2 J_{mm} + \rho\sigma\delta\pi W J_{Wm}\Bigr\}, \end{aligned}$$

where time t subscripts are omitted and the subscripts on J denote partial derivatives (e.g., $\partial J/\partial W = J_W$).

Fist Order Condition:

$$\begin{aligned} C &= (J_W)^{-\psi}\bigl[(1-\gamma)J\bigr]^{(1-\gamma\psi)/(1-\gamma)}\beta^{\psi}, \\ \pi &= -\frac{(m-r)J_W}{\sigma^2 W J_{WW}} - \frac{\rho\delta J_{Wm}}{\sigma W J_{WW}}. \end{aligned}$$

Substituting these first-order conditions into the equation above and rearranging gives the Bellman equation

$$\begin{aligned} 0 = {} & \mathrm{f}(C, J) - \frac{(m-r)^2 J_W^2}{2\sigma^2 J_{WW}} - \frac{\rho\delta(m-r)J_W J_{Wm}}{\sigma J_{WW}} \\ & + rW J_W - (J_W)^{1-\psi}\bigl[(1-\gamma)J\bigr]^{(1-\gamma\psi)/(1-\gamma)}\beta^{\psi} \\ & + \kappa(\theta - m)J_m + \frac{1}{2}\delta^2 J_{mm} - \frac{\rho^2\delta^2 J_{Wm}^2}{2J_{WW}}. \end{aligned}$$

We guess a solution of the form

$$J(W_t, m_t) = I(m_t)\frac{W_t^{1-\gamma}}{1-\gamma}.$$

Substituting this solution into the Bellman equation and simplifying yields the equation:

$$\begin{aligned}0 = &-\frac{1}{1-\psi}\beta^{\psi} I^{1+(1-\psi)/(1-\gamma)} + \frac{\psi}{1-\psi}\beta I + \frac{(m-r)^2}{2\gamma\sigma^2} I + \frac{\rho\delta(m-r)}{\sigma\gamma} I_m + rI \\ &+ \kappa(\theta - m)\frac{1}{1-\gamma} I_m + \frac{1}{2}\frac{\delta^2}{1-\gamma} I_{mm} + \frac{\rho^2\delta^2}{2\gamma}\frac{I_m^2}{I}.\end{aligned}$$

We now make the transformation

$$I = H^{-(1-\gamma)/(1-\psi)}.$$

The derivatives of I are

$$\begin{aligned}I_m &= -\frac{1-\gamma}{1-\psi} H^{-(1-\gamma)/(1-\psi)-1} H_m, \\ I_{mm} &= \frac{1-\gamma}{1-\psi}\left(\frac{1-\gamma}{1-\psi} + 1\right) H^{-(1-\gamma)/(1-\psi)-2} (H_m)^2 \\ &\quad - \frac{1-\gamma}{1-\psi} H^{-(1-\gamma)/(1-\psi)-1} H_{mm}.\end{aligned}$$

Substituting these expressions into the ODE above yields (after some simplification) the following ODE

$$\begin{aligned}0 = &-\frac{1}{1-\psi}\beta^{\psi} H^{-(1-\gamma)/(1-\psi)-1} + \frac{\psi}{1-\psi}\beta H^{-(1-\gamma)/(1-\psi)} \\ &+ \frac{(m-r)^2}{2\gamma\sigma^2} H^{-(1-\gamma)/(1-\psi)} - \frac{\rho\delta(m-r)}{\sigma\gamma}\frac{1-\gamma}{1-\psi} H^{-(1-\gamma)/(1-\psi)-1} H_m \\ &+ rH^{-(1-\gamma)/(1-\psi)} - \kappa(\theta - m)\frac{1}{1-\psi} H^{-(1-\gamma)/(1-\psi)-1} H_m \\ &+ \frac{1}{2}\frac{\delta^2}{1-\gamma}\frac{1-\gamma}{1-\psi}\left(\frac{1-\gamma}{1-\psi} + 1\right) H^{-(1-\gamma)/(1-\psi)-2} (H_m)^2 \\ &- \frac{1}{2}\frac{\delta^2}{1-\gamma}\frac{1-\gamma}{1-\psi} H^{-(1-\gamma)/(1-\psi)-1} H_{mm} \\ &+ \frac{\rho^2\delta^2}{2\gamma}\left(\frac{1-\gamma}{1-\psi} H^{-(1-\gamma)/(1-\psi)-1} H_m\right)^2 H^{(1-\gamma)/(1-\psi)}.\end{aligned}$$

Simplifying this expression results in the following equation

$$
\begin{aligned}
0 = {} & -\beta^{\psi} H^{-1} + \psi\beta + \frac{(1-\psi)}{2\gamma\sigma^2}(m-r)^2 - \frac{\rho\delta(1-\gamma)}{\sigma\gamma}(m-r)\frac{H_m}{H} \\
& + r(1-\psi) - \kappa(\theta - m)\frac{H_m}{H} + \frac{1}{2}\delta^2\left(\frac{1-\gamma}{1-\psi} + 1\right)\left(\frac{H_m}{H}\right)^2 \\
& - \frac{1}{2}\delta^2\frac{H_{mm}}{H} + \frac{\rho^2\delta^2(1-\gamma)^2}{2\gamma(1-\psi)}\left(\frac{H_m}{H}\right)^2.
\end{aligned}
$$

This equation is a non-homogeneous ODE. It can be shown that the general solution to the associated homogeneous equation belongs to the degenerate hypergeometric equation class. However, this solution is as hard to interpret as the ODE equation itself and must be solved using numerical approximations.

It is still possible to find an approximate analytical solution to this ODE that gives us a much deeper understanding of the problem. This approximate solution is based on a log-linear expansion of the consumption–wealth ratio around its unconditional mean. This is exactly the same type of approximation that Campbell (1993) and Campbell and Viceira (1999, 2001) pursue in discrete time. However, while they use the approximation to linearize the log budget constraint, we use it to solve the Bellman equation.

As is shown in Proposition 6 above, we let

$$
\begin{aligned}
\beta^{\psi} H_t^{-1} &\approx \exp\{\overline{c-w}\} + \exp\{\overline{c-w}\}\big[(c_t - w_t) - (\overline{c-w})\big] \\
&\equiv h_0 + h_1(c_t - w_t),
\end{aligned}
$$

where $(\overline{c-w}) = \mathrm{E}[c_t - w_t]$, $h_0 = \exp\{\overline{c-w}\}[1 - (\overline{c-w})]$ and $h_1 = \exp\{\overline{c-w}\}$.

Substituting the linear approximation for $\beta^{\psi} H_t^{-1}$ in the first term of the Bellman equation, we see that the resulting ODE has a solution of the form $H_t = \exp[Am_t^2 + Bm_t + D]$. This solution implies that $c_t - x_t = -Am_t^2 - Bm_t - D + \psi\log\beta$.

Therefore, our approach replaces the term that causes the non-linear ODE to be non-solvable analytically with a log-linear approximation that transforms the equation into another ODE with known analytical solution; if the log-linear approximation is accurate, the exact analytical solution to the approximate ODE will also verify the original ODE subject to some approximation error. In this sense it is an approximate analytical solution. Nevertheless, we show further below that the approximation error is zero for the special cases of log utility and constant investment opportunities.

Substituting the solution form $H_t = \exp[Am_t^2 + Bm_t + D]$ into this approximate ODE leads to the equation

$$
\begin{aligned}
0 = {} & -h_0 - h_1\big(-Am_t^2 - Bm_t - D + \psi\log\beta\big) + \psi\beta + r(1-\psi) \\
& + \frac{1-\psi}{2\gamma\sigma^2}\big(m_t^2 - 2rm_t + r^2\big) - \frac{\rho\delta(1-\gamma)}{\sigma\gamma}\big(2Am_t^2 + Bm_t - 2rAm_t + rB\big)
\end{aligned}
$$

$$- 2\kappa\theta A m_t - \kappa\theta B + 2\kappa A m_t^2 + \kappa B m_t + a\big(4A^2 m_t^2 + 4AB m_t + B^2\big)$$
$$- \delta^2 A - 2\delta^2 A^2 m_t^2 - 2\delta^2 AB m_t - \frac{1}{2}\delta^2 B^2$$

where

$$a = \frac{1}{2}\delta^2(1-\gamma)\left(\frac{1}{1-\psi} + \frac{1}{1-\gamma} + \frac{\rho^2(1-\gamma)}{\gamma(1-\psi)}\right).$$

Separating variables leads to a system of three algebraic equations for A, B, and D, which can easily be solved in closed-form

$$0 = \big(4a - 2\delta^2\big)A^2 + \left[2\kappa + h_1 - \frac{2\rho\delta(1-\gamma)}{\sigma\gamma}\right]A + \frac{1-\psi}{2\gamma\sigma^2},$$
$$0 = \left[4aA - 2\delta^2 A + \kappa - h_1 - \frac{\rho\delta(1-\gamma)}{\sigma\gamma}\right]B + \frac{2r\rho\delta(1-\gamma)}{\sigma\gamma}A - 2\kappa\theta A - \frac{r(1-\psi)}{\gamma\sigma^2},$$
$$0 = h_1 D - \delta^2 A^2 + \left[a - \frac{1}{2}\delta^2\right]B^2 - \left[\kappa\theta - \frac{\rho\delta r(1-\gamma)}{\sigma\gamma}\right]B - h_0 + \psi\beta + \frac{r^2(1-\psi)}{2\gamma\sigma^2} + r(1-\psi) - h_1\psi\log\beta.$$

Therefore, the approximate value function of the problem is

$$J(W_t, m_t) = \exp\left\{-\frac{1-\gamma}{1-\psi}\big[Am_t^2 + Bm_t + D\big]\right\}\frac{W_t^{1-\gamma}}{1-\gamma},$$

where A, B, and C are solutions to the system of equations above.

Substituting the solution for the indirect utility function given in Proposition 6 into the first order conditions, we can solve for the optimal policies

$$\frac{C_t}{W_t} = \beta^{\psi}\exp\big[-Am_t^2 - Bm_t - D\big], \quad \text{and}$$
$$\pi_t = \frac{m-r}{\gamma\sigma^2} + \frac{\rho\delta}{\gamma\sigma}\frac{1-\gamma}{1-\psi}(2Am_t + B).$$

5. Example

We conclude with two other examples to illustrate the techniques discussed above. The first example is the case of time-varying volatility and the second example deals with time-varying interest rates.

5.1. Time-varying volatility

Chacko and Viceira (2005) studied this case and solved the model using a perturbation around the long-run consumption-to-wealth ratio. Here, we show how to solve the model exactly in two cases: first when the elasticity of substitution parameter equals 1; and second in the case of a finite horizon and terminal utility of wealth. We then apply the two perturbation methods we have discussed to generate approximate results and check the accuracy of these results by comparing them to the two cases where we have exact solutions.

For this example we assume that we have constant riskless interest rates and constant expected returns for risky assets. However, the volatility of the risky assets is time-varying. So the stock dynamics is

$$\begin{aligned}
&\frac{\mathrm{d}S_t}{S_t} = m\,\mathrm{d}t + \sqrt{v_t}\,\mathrm{d}Z_{S,t},\\
&\frac{\mathrm{d}B_t}{B_t} = r\,\mathrm{d}t,\\
&\mathrm{d}\left(\frac{1}{v_t}\right) = \mathrm{d}y_t = \kappa(\theta - y_t)\,\mathrm{d}t + \sigma\sqrt{y_t}\,\mathrm{d}Z_{y,t}
\end{aligned} \tag{21}$$

where the Wiener processes $\mathrm{d}Z_{S,t}$ and $\mathrm{d}Z_{v,t}$ are correlated by $\mathrm{Corr}(\mathrm{d}Z_{S,t}, \mathrm{d}Z_{y,t}) = \rho$. Instead of considering the instantaneous variance v_t, we usually consider its reciprocal $y_t = 1/v_t$ to be a state variable and thus affine in our example.

Using Itô's lemma, we can easily derive the dynamics of v_t from (21). It follows a mean-reverting, square root process

$$\frac{\mathrm{d}v_t}{v_t} = \kappa_v(\theta_v - v_t) - \sigma\sqrt{v_t}\,\mathrm{d}Z_{y,t}$$

where $\theta_v = (\theta - \sigma^2/\kappa)^{-1}, \kappa_v = \kappa/\theta_v$.

We also assume an infinite horizon for this example, so the recursive utility function is

$$J_t = \sup_{\pi,C} E_t\left\{\int_t^\infty \frac{\beta(1-\gamma)J_t}{1-1/\psi}\left[\left(\frac{C_t}{((1-\gamma)J_t)^{1/(1-\gamma)}}\right)^{1-1/\psi} - 1\right]\mathrm{d}t\right\}.$$

From the Principle of Optimality, the indirect utility function should satisfy

$$\begin{aligned}
0 = \sup_{\pi,C}\Bigg\{&\frac{\beta(1-\gamma)J}{1-1/\psi}\left[\left(\frac{C}{((1-\gamma)J)^{1/(1-\gamma)}}\right)^{1-1/\psi} - 1\right]\\
&+ \big[\pi(m-r)W + rW - C\big]J_W + \frac{1}{2y}\pi^2 W^2 J_{WW} + \kappa(\theta - y)J_y\\
&+ \frac{1}{2}\sigma^2 y J_{yy} + \rho\pi\sigma W J_{yW}\Bigg\}
\end{aligned} \tag{22}$$

where the time subscripts t are omitted.

The first-order condition is

$$C_t = (J_W)^{-\psi}\big[(1-\gamma)J\big]^{(1-\gamma\psi)/(1-\gamma)}\beta^{\psi},$$
$$\pi_t = -\frac{y}{WJ_{WW}}\big((m-r)J_W + \rho\sigma J_{yW}\big).$$

Substituting the FOC into (22), we yield the Bellmen equation

$$\begin{aligned}0 = {} & \frac{\beta(1-\gamma)}{1-1/\psi}J\left[\left(\frac{(J_W)^{-\psi}[(1-\gamma)J]^{(1-\gamma\psi)/(1-\gamma)}\beta^{\psi}}{((1-\gamma)J)^{1/(1-\gamma)}}\right)^{1-1/\psi} - 1\right] \\ & - (J_W)^{1-\psi}\big[(1-\gamma)J\big]^{(1-\gamma\psi)/(1-\gamma)}\beta^{\psi} + \kappa(\theta-y)J_y + \frac{1}{2}\sigma^2 y J_{yy} \\ & + rWJ_W - \frac{1}{2}\frac{y}{J_{WW}}\big((m-r)J_W + \rho\sigma J_{yW}\big)^2.\end{aligned}$$

If we guess the solution form as $J_t = I(y_t)\frac{W_t^{1-\gamma}}{1-\gamma}$ and substitute it into the Bellman equation, we get the following ODE

$$\begin{aligned}0 = {} & \frac{\beta^{\psi}}{\psi-1}I^{1+(1-\psi)/(1-\gamma)} - \frac{\psi}{\psi-1}\beta I + \frac{(m-r)^2}{2\gamma}yI + \frac{\rho\sigma y(m-r)I_y}{\gamma} \\ & + \frac{\rho^2\sigma^2 y I_y^2}{2\gamma I} + \frac{\kappa(\theta-y)I_y}{1-\gamma} + \frac{\sigma^2 y I_{yy}}{2(1-\gamma)} + rI, \end{aligned} \tag{23}$$

where again, all subscripts t are omitted.

5.1.1. Exact solution when $\psi = 1$

When $\psi \to 1$, Equation (23) can be simplified to

$$\begin{aligned}0 = {} & \beta\log\beta I - \frac{\beta}{1-\gamma}I\log I - \beta I + \frac{(m-r)^2}{2\gamma}yI + \frac{\rho\sigma y(m-r)I_y}{\gamma} \\ & + \frac{\rho^2\sigma^2 y I_y^2}{2\gamma I} + \frac{\kappa(\theta-y)I_y}{1-\gamma} + \frac{\sigma^2 y I_{yy}}{2(1-\gamma)} + rI. \end{aligned} \tag{24}$$

It is easy to see that this ODE has a solution for I in the form $I = \exp(Ay_t + D)$. Plugging this into (24), we get the equation for A and D

$$\begin{aligned}0 &= \left[\frac{\rho^2}{\gamma} + \frac{1}{1-\gamma}\right]\frac{\sigma^2}{2}A^2 - \left[\frac{\beta+\kappa}{1-\gamma} - \frac{\rho\sigma(m-r)}{\gamma}\right]A + \frac{(m-r)^2}{2\gamma}, \\ 0 &= (1-\gamma)(\beta\log\beta - \beta + r) - \beta D + \kappa\theta A. \end{aligned} \tag{25}$$

Therefore, we successfully reduce the problem to a quadratic Equation (25). Now given that we have the suitable solutions for (24), we can show the corresponding optimal asset allocation ratio is

$$\frac{C_t}{W_t} = \beta,$$

$$\pi_t = \frac{1}{\gamma v_t}(m - r + \rho\sigma A).$$

This result is intuitive. It shows that an individual is under the unit elasticity assumption indifferent to the allocation of consumption over time. Therefore, only the time preference parameter β can affect the consumption/investment ratio of an individual.

5.1.2. Approximation methods when $\psi \neq 1$

However, similar to Proposition 4, if the unit elasticity assumption does not apply, we do not have the analytic solution for (23). In this case we have to apply the approximation methods shown above.

5.1.2.1. Perturbation method 1 (around unit elasticity). If we think the elasticity ψ is around 1, then we can linearize the first term in (23)

$$I^{1+(1-\psi)/(1-\gamma)} \approx I + \frac{1-\psi}{1-\gamma} I \log I.$$

The result is an analytically solvable equation

$$\begin{aligned} 0 = {} & \frac{\beta^\psi - \psi\beta}{\psi - 1} I - \frac{\beta^\psi}{1-\gamma} I \log I + \frac{(m-r)^2}{2\gamma} yI + \frac{\rho\sigma y(m-r) I_y}{\gamma} \\ & + \frac{\rho^2\sigma^2 y I_y^2}{2\gamma I} + \frac{\kappa(\theta - y) I_y}{1-\gamma} + \frac{\sigma^2 y I_{yy}}{2(1-\gamma)} + rI. \end{aligned} \tag{26}$$

According to Proposition 4, we know that this equation has an analytic solution of I in the form of $I = \exp[Ay + By^2 + D]$. Plugging this in, we get the equations for A, B and D:

$$\begin{aligned} 0 = {} & \frac{\beta^\psi - \psi\beta}{\psi - 1} - \frac{\beta^\psi}{1-\gamma}(Ay + By^2 + D) + \frac{\rho^2\sigma^2 y(2By + A)^2}{2\gamma} \\ & + \frac{(m-r)^2}{2\gamma} y + \frac{\rho\sigma y(m-r)(2By + A)}{\gamma} + \frac{\kappa(\theta - y)(2By + A)}{1-\gamma} \\ & + \frac{\sigma^2 y(4B^2y^2 + 4ABy + A^2 + 2B)}{2(1-\gamma)} + r. \end{aligned} \tag{27}$$

This is solvable iff $B = 0$ since the coefficient of y^3 is

$$0 = \frac{2\rho^2\sigma^2 B^2}{\gamma} + \frac{2\sigma^2 B^2}{(1-\gamma)}.$$

Equation (27) can now be simplified to

$$0 = \frac{\beta^\psi - \psi\beta}{\psi - 1} - \frac{\beta^\psi}{1-\gamma}(Ay + D) + \frac{(m-r)^2}{2\gamma} y + r$$

$$
\begin{aligned}
&+\frac{\rho\sigma y(m-r)A}{\gamma}+\frac{\kappa A(\theta-y)}{1-\gamma}+\frac{\rho^2\sigma^2 yA^2}{2\gamma}+\frac{\sigma^2 yA^2}{2(1-\gamma)},\\
0=&\frac{(1-\gamma)(m-r)^2}{2}+\big[\rho\sigma(m-r)(1-\gamma)-\gamma\big(\kappa+\beta^{\psi}\big)\big]A\\
&+\big[\rho^2(1-\gamma)+\gamma\big]\frac{\sigma^2}{2}A^2,\\
0=&\frac{(m-r)^2}{2\gamma}+\left[\frac{\rho\sigma(m-r)}{\gamma}-\frac{\kappa+\beta^{\psi}}{1-\gamma}\right]A+\left[\frac{\rho^2}{\gamma}+\frac{1}{1-\gamma}\right]\frac{\sigma^2}{2}A^2,\\
0=&\frac{\beta^{\psi}-\psi\beta}{\psi-1}+\frac{\kappa\theta A-D\beta^{\psi}}{1-\gamma}+r.
\end{aligned}
$$

After we solve for A and D, we know that

$$J_t=\exp(Ay_t+D)\frac{W_t^{1-\gamma}}{1-\gamma}.$$

The optimal consumption/investment allocation is

$$
\begin{aligned}
\frac{C_t}{W_t}&=\beta^{\psi}\exp\left[\left(-\frac{A}{v_t}-D\right)\frac{1-\psi}{1-\gamma}\right],\\
\pi_t&=\frac{1}{\gamma v_t}(m-r+\rho\sigma A).
\end{aligned}
$$

5.1.2.2. Perturbation method 2 (around mean consumption/wealth ratio). We can further simplify the equation by assuming $I=H^{-(1-\gamma)/(1-\psi)}$, i.e.,

$$
\begin{aligned}
0=&-\beta^{\psi}H^{-1}+\psi\beta+\frac{(1-\psi)(m-r)^2}{2\gamma}y-\frac{\rho\sigma(m-r)(1-\gamma)}{\gamma}\frac{H_y}{H}y\\
&+\frac{\rho^2\sigma^2(1-\gamma)^2}{2\gamma(1-\psi)}\left(\frac{H_y}{H}\right)^2y+\frac{(2-\gamma-\psi)\sigma^2}{2(1-\psi)}\left(\frac{H_y}{H}\right)^2y\\
&-\kappa(\theta-y)\frac{H_y}{H}-\frac{\sigma^2}{2}\frac{H_{yy}}{H}y+r(1-\psi).
\end{aligned}
\tag{28}
$$

Since in this example we assume m and r to be constant, it is easy to see that the dynamics of volatility can be captured entirely by the dynamics of the Sharpe ratio X. Similar to 5 and to our procedure above, we make the following approximation to render (28) solvable

$$\beta^{\psi}H^{-1}=\exp(c_t-w_t)\approx h_0+h_1(c_t-w_t).\tag{29}$$

Substituting this into (28), yields a solvable ODE with a solution in the form $H=\exp\{Ay_t+D\}$. If we substitute it in the approximated version of (28), we get

$$0=-h_0-h_1(\psi\log\beta-Ay_t-D)+\psi\beta+\frac{(1-\psi)(m-r)^2}{2\gamma}y$$

$$- \frac{\rho\sigma(m-r)(1-\gamma)}{\gamma} yA + \frac{\rho^2\sigma^2(1-\gamma)^2}{2\gamma(1-\psi)} A^2 y$$
$$+ \frac{(1-\gamma)\sigma^2}{2(1-\psi)} A^2 y - \kappa(\theta - y)A + r(1-\psi)$$

which leads to

$$0 = \frac{\sigma^2(1-\gamma)}{2\gamma(1-\psi)}\left[\rho^2(1-\gamma)+\gamma\right]A^2 + \left[(h_1+\kappa) - \frac{\rho\sigma(1-\gamma)(m-r)}{\gamma}\right]A$$
$$+ \frac{(1-\psi)(m-r)^2}{2\gamma},$$
$$0 = -h_0 + h_1(D - \psi\log\beta) + \psi\beta - \kappa\theta A + r(1-\psi). \tag{30}$$

So the approximate indirect utility function is

$$J(W_t, v_t) = \exp\left[-\frac{1-\gamma}{1-\psi}\left(\frac{A}{v_t} + B\right)\right]\frac{W_t^{1-\gamma}}{1-\gamma}.$$

We know the optimal consumption ratio and portfolio allocation immediately

$$\frac{C_t}{W_t} = \beta^{\psi}\exp\left(-\frac{A}{v_t} - D\right),$$
$$\pi_t = \frac{1}{\gamma v_t}\left(m - r - \frac{1-\gamma}{1-\psi}\rho\sigma A\right).$$

5.1.3. Numerical results

Now that we have the formulas, we would like to apply them in order to test the accuracy of the approximations. We use the parameter estimates from Chacko and Viceira (2005), which we give in Table 1.

Table 2 presents results from the case where the elasticity of substitution equals one and compares them to the results when the elasticity of substitution parameter is close to one. In looking at the mean allocation to stocks, the ratio of hedging to myopic demand, and the long-term expected return on wealth, we see that the approximate results (generated from the perturbation methods) are very close to the exact solution, as they should be. This indicates that the approximations are accurate.

Tables 3 and 4 present further results for a wider range of risk aversion parameters and elasticities of substitution. What we see from this table is that the further the elasticity of substitution is from one, the greater the disparity in the results between the two perturbation methods. Essentially, the method which expands around the consumption to wealth ratio produces a more accurate result for elasticities of substitution very different from one. However, for elasticities of substitution close to one, the results for the two methods are nearly identical. What we can conclude, therefore, is that in cases where a closed-form solution is important and one is only concerned with elasticities

Table 1
Estimates of the stochastic process for returns and volatility. Model: Time varying volatility (Section 5.1). Parameter estimates (monthly)

	1926.01 – 2000.12
$\mu - r$	0.0811
κ	0.3374
θ	27.9345
σ	0.6503
ρ	0.5241
r	0.0150
β	0.0600

Note: This table reports spectral GMM estimates of the stochastic process driving stock returns and volatility. The estimation is described in detail in Chacko and Viceira (2005). The monthly estimates are based on excess stock returns on the CRSP value-weighted portfolio over the T-bill rate from January 1926 through December 2000.

of substitution equal or close to one, the perturbation method that expands around the elasticity of substitution offers a very compelling solution method. However, in cases where a closed-form solution is not critical and one might be dealing with elasticities of substitution very different from one, the perturbation method that expands around the consumption to wealth ratio is the more appropriate choice.

5.1.4. Limiting behavior of the approximated results

Our approximation method is a generalization of the exact unit elasticity case. Therefore, the solution of both methods should converge to the exact solution when $\psi \to 1$.

5.1.4.1. Perturbation method 1. The equation for A and D is

$$
\begin{aligned}
0 &= \frac{(1-\gamma)(m-r)^2}{2} + \left[\rho\sigma(m-r)(1-\gamma) - \gamma\left(\kappa + \beta^{\psi}\right)\right]A \\
&\quad + \left[\rho^2(1-\gamma) + \gamma\right]\frac{\sigma^2}{2}A^2, \\
0 &= \frac{(m-r)^2}{2\gamma} + \left[\frac{\rho\sigma(m-r)}{\gamma} - \frac{\kappa + \beta^{\psi}}{1-\gamma}\right]A + \left[\frac{\rho^2}{\gamma} + \frac{1}{1-\gamma}\right]\frac{\sigma^2}{2}A^2, \\
0 &= \frac{\beta^{\psi} - \psi\beta}{\psi - 1} + \frac{\kappa\theta A - D\beta^{\psi}}{1-\gamma} + r
\end{aligned}
$$

Table 2
Comparison of the exact solution with results of perturbation methods 1 and 2. Mean optimal percentage allocation to stocks, percentage hedging demand over myopic demand and long-term expected return on wealth (sample: 1926.01 – 2000.12)

R.R.A.	Exact solution ($\psi = 1$)	Perturbation method 1 ($\psi = 1.1$)	Perturbation method 2 ($\psi = 1.1$)
(A) Mean optimal allocation to stocks (%): $E[\pi t(yt)] = \pi(\theta) \times 100$			
0.75	305.66	305.81	305.74
1.00	226.55	226.55	226.55
1.50	149.32	149.25	149.32
2.00	111.37	111.30	111.37
4.00	55.24	55.18	55.25
10.00	21.99	21.96	22.00
20.00	10.98	10.96	10.98
(B) Ratio of hedging demand over myopic demand (%):			
0.75	1.19	1.24	1.22
1.00	0.00	0.00	0.00
1.50	−1.13	−1.18	−1.14
2.00	−1.68	−1.75	−1.68
4.00	−2.47	−2.57	−2.46
10.00	−2.94	−3.05	−2.91
20.00	−3.09	−3.21	−3.06
(C) Long-term expected return on wealth (%): $(\pi(\theta)(\mu - r) + r) \times 100$			
0.75	26.29	26.30	26.30
1.00	19.87	19.87	19.87
1.50	13.61	13.60	13.61
2.00	10.53	10.53	10.53
4.00	5.98	5.98	5.98
10.00	3.28	3.28	3.28
20.00	2.39	2.39	2.39

Note: Panel A reports mean optimal percentage allocations to stocks for different coefficients of relative risk aversion (R.R.A.) and for elasticities of intertemporal substitution of consumption (ψ) equal to 1 in the context of the exact solution and equal to 1.1 in the context of perturbation methods 1 and 2. Panel B reports the percentage ratio of intertemporal hedging portfolio demand over myopic portfolio demand, which is independent of the level of precision or volatility. Panel C reports the percentage unconditional mean of the log return on wealth. These numbers are based on the monthly parameter estimates of the joint process for return and volatility reported in Table 1.

apparently, when $\psi \to 1$. Then these equations become Equations (25) for the unit elasticity case since

$$\lim_{\psi \to 1} \frac{\beta^{\psi} - \psi\beta}{\psi - 1} = \beta \log \beta - \beta.$$

Table 3
Perturbation method 1: Perturbation around unit elasticity of substitution. Mean optimal percentage allocation to stocks, percentage hedging demand over myopic demand and long-term expected return on wealth (sample: 1926.01 – 2000.12)

R.R.A.								
(A) Mean optimal allocation to stocks (%): $E[\pi t(yt)] = \pi(\theta) \times 100$								
0.75	306.03	305.66	304.96	304.49	303.76	303.35	303.23	
1.00	226.55	226.55	226.55	226.55	226.55	226.55	226.55	
1.50	149.16	149.32	149.64	149.86	150.21	150.40	150.46	
2.00	111.19	111.37	111.72	111.96	112.35	112.57	112.63	
4.00	55.10	55.24	55.49	55.66	55.95	56.11	56.16	
10.00	21.93	21.99	22.11	22.19	22.33	22.40	22.43	
20.00	10.94	10.98	11.04	11.08	11.15	11.19	11.21	
	1/0.75	1.00	1/1.5	1/2	1/4	1/10	1/20	E.I.S
(B) Ratio of hedging demand over myopic demand (%):								
0.75	1.31	1.19	0.96	0.80	0.56	0.43	0.38	
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
1.50	−1.24	−1.13	−0.92	−0.78	−0.55	−0.42	−0.38	
2.00	−1.84	−1.68	−1.37	−1.16	−0.82	−0.62	−0.57	
4.00	−2.71	−2.47	−2.02	−1.72	−1.21	−0.93	−0.84	
10.00	−3.21	−2.94	−2.41	−2.05	−1.45	−1.11	−1.01	
20.00	−3.38	−3.09	−2.54	−2.15	−1.53	−1.17	−1.07	
	1/0.75	1.00	1/1.5	1/2	1/4	1/10	1/20	E.I.S
(C) Long-term expected return on wealth (%): $(\pi(\theta)(\mu - r) + r) \times 100$								
0.75	26.32	26.29	26.23	26.19	26.13	26.10	26.09	
1.00	19.87	19.87	19.87	19.87	19.87	19.87	19.87	
1.50	13.60	13.61	13.64	13.65	13.68	13.70	13.70	
2.00	10.52	10.53	10.56	10.58	10.61	10.63	10.63	
4.00	5.97	5.98	6.00	6.01	6.04	6.05	6.05	
10.00	3.28	3.28	3.29	3.30	3.31	3.32	3.32	
20.00	2.39	2.39	2.40	2.40	2.40	2.41	2.41	
	1/0.75	1.00	1/1.5	1/2	1/4	1/10	1/20	E.I.S

Note: Panel A reports mean optimal percentage allocations to stocks for different coefficients of relative risk aversion (R.R.A.) and for elasticities of intertemporal substitution of consumption (E.I.S.). Panel B reports the percentage ratio of intertemporal hedging portfolio demand over myopic portfolio demand, which is independent of the level of precision or volatility. Panel C reports the percentage unconditional mean of the log return on wealth. These numbers are based on the monthly parameter estimates of the joint process for return and volatility reported in Table 1.

So $\frac{C_t}{W_t}$ and π_t converge to the unit elasticity case correspondingly

$$\lim_{\psi \to 1} \frac{C_t}{W_t} = \lim_{\psi \to 1} \beta^{\psi} \exp\left[\left(-\frac{A}{v_t} - D\right)\frac{1-\psi}{1-\gamma}\right] = \beta, \quad \pi_t = \frac{1}{\gamma v_t}(m - r + \rho\sigma A).$$

Table 4
Perturbation method 2: Perturbation around mean of consumption/wealth ratio. Mean optimal percentage allocation to stocks, percentage hedging demand over myopic demand and long-term expected return on wealth (sample: 1926.01 – 2000.12)

R.R.A.								
(A) Mean optimal allocation to stocks (%): $E[\pi t(yt)] = \pi(\theta) \times 100$								
0.75	305.92	305.66	305.42	305.32	305.17	305.09	305.07	
1.00	226.55	226.55	226.55	226.55	226.55	226.55	226.55	
1.50	149.30	149.32	149.34	149.35	149.37	149.38	149.38	
2.00	111.38	111.37	111.37	111.37	111.37	111.37	111.37	
4.00	55.26	55.24	55.21	55.20	55.18	55.16	55.16	
10.00	22.01	21.99	21.97	21.96	21.94	21.93	21.93	
20.00	10.99	10.98	10.97	10.96	10.95	10.94	10.94	
	1/0.75	1.00	1/1.5	1/2	1/4	1/10	1/20	E.I.S.
(B) Ratio of hedging demand over myopic demand (%):								
0.75	1.28	1.19	1.11	1.08	1.03	1.00	0.99	
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
1.50	−1.15	−1.13	−1.12	−1.11	−1.10	−1.10	−1.09	
2.00	−1.68	−1.68	−1.68	−1.68	−1.68	−1.68	−1.68	
4.00	−2.43	−2.47	−2.52	−2.54	−2.58	−2.60	−2.61	
10.00	−2.86	−2.94	−3.02	−3.07	−3.14	−3.18	−3.20	
20.00	−3.00	−3.09	−3.19	−3.25	−3.33	−3.38	−3.40	
	1/0.75	1.00	1/1.5	1/2	1/4	1/10	1/20	E.I.S.
(C) Long-term expected return on wealth (%): $(\pi(\theta)(\mu - r) + r) \times 100$								
0.75	26.31	26.29	26.27	26.26	26.25	26.24	26.24	
1.00	19.87	19.87	19.87	19.87	19.87	19.87	19.87	
1.50	13.61	13.61	13.61	13.61	13.61	13.61	13.61	
2.00	10.53	10.53	10.53	10.53	10.53	10.53	10.53	
4.00	5.98	5.98	5.98	5.98	5.97	5.97	5.97	
10.00	3.28	3.28	3.28	3.28	3.28	3.28	3.28	
20.00	2.39	2.39	2.39	2.39	2.39	2.39	2.39	
	1/0.75	1.00	1/1.5	1/2	1/4	1/10	1/20	E.I.S.

Note: Panel A reports mean optimal percentage allocations to stocks for different coefficients of relative risk aversion (R.R.A.) and for elasticities of intertemporal substitution of consumption (E.I.S.). Panel B reports the percentage ratio of intertemporal hedging portfolio demand over myopic portfolio demand, which is independent of the level of precision or volatility. Panel C reports the percentage unconditional mean of the log return on wealth. These numbers are based on the monthly parameter estimates of the joint process for return and volatility reported in Table 1.

5.1.4.2. Perturbation method 2. The equation for A and D is

$$\begin{aligned}0 &= \frac{\sigma^2(1-\gamma)}{2\gamma(1-\psi)}\big[\rho^2(1-\gamma)+\gamma\big]A^2 \\ &\quad + \left[(h_1+\kappa) - \frac{\rho\sigma(1-\gamma)(m-r)}{\gamma}\right]A + \frac{(1-\psi)(m-r)^2}{2\gamma}, \\ 0 &= -h_0 + h_1(D - \psi\log\beta) + \psi\beta - \kappa\theta A + r(1-\psi). \end{aligned} \tag{31}$$

The first quadratic equation can be rewritten as

$$\begin{aligned}0 &= \frac{\sigma^2}{2\gamma}\big[\rho^2(1-\gamma)+\gamma\big]\left[\frac{(1-\gamma)A}{1-\psi}\right]^2 \\ &\quad + \left[\frac{h_1+\kappa}{1-\gamma} - \frac{\rho\sigma(m-r)}{\gamma}\right]\frac{(1-\gamma)A}{1-\psi} + \frac{(1-\gamma)(m-r)^2}{2\gamma}. \end{aligned} \tag{32}$$

Comparing (32) to (25), we can see

$$\lim_{\psi\to 1}\frac{(1-\gamma)A}{1-\psi} = A^* \quad \text{if } \lim_{\psi\to 1} h_1 = \beta$$

where A^* denotes the exact solution of A in the unit elasticity case.

So let us establish that $\lim_{\psi\to 1} h_1 = \beta$. This is true because when $\psi \to 1$, the first quadratic equation can be written as

$$0 = \frac{\sigma^2(1-\gamma)}{2\gamma}\big[\rho^2(1-\gamma)+\gamma\big]A^2,$$

i.e., $\lim_{\psi\to 1} A = 0$. In addition, as $\psi \to 1$ the rest of the equations become

$$\begin{aligned} h_0 &= h_1(1-\log h_1), \\ h_1 &= \frac{C_t}{W_t} = \beta\exp(-D), \\ 0 &= -h_0 + h_1(D - \log\beta) + \beta - \kappa\theta A. \end{aligned}$$

The solution is $h_1 = \beta$, $D = 0$.

5.1.5. Exact solution for terminal utility of wealth

In all previous subsections we are discussing the infinite horizon case. However, in this subsection, we will illustrate our technique under the finite horizon assumption.

By the Principle of Optimality (6), the optimal allocation should satisfy

$$\begin{aligned}0 &= \sup_{\pi,C}\Big\{u(C) + \big[\pi(m-r)W + rW - C\big]J_W + \frac{1}{2y}\pi^2W^2J_{WW} \\ &\quad + \kappa(\theta - y)J_y + \frac{1}{2}\sigma^2 y J_{yy} + \rho\pi\sigma W J_{yW} - J_\tau\Big\} \end{aligned} \tag{33}$$

where $\tau = T - t$ is the time to horizon. If we further assume that there is only terminal consumption, i.e., $C_t = 0$ for $t \neq T$, then (33) can be simplified to

$$0 = \sup_{\pi, C}\Big\{\big[\pi(m-r)+r\big]WJ_W + \frac{1}{2y}\pi^2 W^2 J_{WW} + \kappa(\theta - y)J_y + \frac{1}{2}\sigma^2 y J_{yy} + \rho\pi\sigma W J_{yW} - J_\tau\Big\}.$$

The FOC can provide us with the optimum allocation between the risk-free and risky assets

$$\pi_t = -\big((m-r)J_W + \rho\sigma J_{yW}\big)\frac{y}{WJ_{WW}}.$$

Substituting this yields the Bellman equation

$$0 = -\frac{y}{2J_{WW}}\big((m-r)J_W + \rho\sigma J_{yW}\big)^2 + rWJ_W + \kappa(\theta - y)J_y + \frac{1}{2}\sigma^2 y J_{yy} - J_\tau.$$

If we guess that $J = I\frac{W^{1-\gamma}}{1-\gamma}$, then

$$0 = \frac{y}{2\gamma}\big((m-r)I + \rho\sigma I_y\big)^2 + rI + \frac{\kappa(\theta - y)}{1-\gamma}I_y + \frac{\sigma^2 y I_{yy}}{2(1-\gamma)} - \frac{I_\tau}{1-\gamma}.$$

As in the previous sections, we guess $I = \exp[A(\tau)y + B(\tau)]$, and then we get the ODE for A and B

$$\frac{\mathrm{d}A(\tau)}{\mathrm{d}\tau} = \frac{1-\gamma}{2\gamma}\big((m-r) + \rho\sigma A(\tau)\big)^2 - \kappa A(\tau) + \frac{1}{2}\sigma^2 A^2(\tau), \tag{34}$$

$$\frac{\mathrm{d}B(\tau)}{\mathrm{d}\tau} = r(1-\gamma) + \kappa\theta A(\tau). \tag{35}$$

Equation (34) can be written as

$$\frac{\mathrm{d}[cA(\tau)]}{\mathrm{d}\tau} = \big(cA(\tau) + r_1\big)\big(cA(\tau) + r_2\big)$$

where

$$c = \frac{\sigma^2}{2\gamma}\big((1-\gamma)\rho^2 + \gamma\big),$$

$$r_1 = \frac{(m-r)(1-\gamma)\rho\sigma - \kappa\gamma}{2\gamma} + \frac{1}{2}\sqrt{\kappa^2 - \frac{1-\gamma}{\gamma}\sigma(m-r)\big(2\rho\kappa + \sigma(m-r)\big)},$$

$$r_2 = \frac{(m-r)(1-\gamma)\rho\sigma - \kappa\gamma}{2\gamma} - \frac{1}{2}\sqrt{\kappa^2 - \frac{1-\gamma}{\gamma}\sigma(m-r)\big(2\rho\kappa + \sigma(m-r)\big)}.$$

So the solution for $A(\tau)$ should satisfy

$$t = \int_0^t \mathrm{d}\tau = \int_0^{cA(t)} \frac{\mathrm{d}\phi}{(\phi + r_1)(\phi + r_2)} = \left[\frac{1}{r_1 - r_2}\log\left|\frac{\phi + r_2}{\phi + r_1}\right|\right]_0^{cA(t)}.$$

The final solutions for A and B are

$$A(\tau) = \frac{2r_1 r_2}{\sigma^2} \frac{\gamma}{((1-\gamma)\rho^2+\gamma)} \frac{\mathrm{e}^{r_1\tau} - \mathrm{e}^{r_2\tau}}{r_1 \mathrm{e}^{r_2\tau} - r_2 \mathrm{e}^{r_1\tau}},$$
$$B(\tau) = r(1-\gamma)\tau + \frac{2\kappa\theta}{\sigma^2} \frac{\gamma}{((1-\gamma)\rho^2+\gamma)} \log\left|\frac{r_2 - r_1}{r_1 \mathrm{e}^{r_1\tau} - r_2 \mathrm{e}^{r_2\tau}}\right|.$$

5.2. *Time-varying interest rates*

In this example, we are dealing with the model that has fixed expected stock returns with a fixed volatility but time-varying interest rates. Expressed in stochastic equations, that is

$$\begin{aligned} \frac{\mathrm{d}S_t}{S_t} &= m\,\mathrm{d}t + \sigma\,\mathrm{d}Z_{S,t}, \\ \frac{\mathrm{d}B_t}{B_t} &= r_t\,\mathrm{d}t, \\ \mathrm{d}r_t &= \kappa(\theta - r_t)\,\mathrm{d}t + \delta\,\mathrm{d}Z_{r,t} \end{aligned} \tag{36}$$

where the Wiener processes $\mathrm{d}Z_{S,t}$ and $\mathrm{d}Z_{r,t}$ are correlated with $\mathrm{Corr}(\mathrm{d}Z_{S,t}, \mathrm{d}Z_{r,t}) = \rho$.

We also assume an infinite horizon for this example, so the recursive utility function is

$$J_t = \sup_{\pi,C} E_t\left\{\int_t^\infty \frac{\beta(1-\gamma)J_s}{1-1/\psi}\left[\left(\frac{C_s}{((1-\gamma)J_s)^{1/(1-\gamma)}}\right)^{1-1/\psi} - 1\right]\mathrm{d}s\right\}.$$

From the Principle of Optimality, the indirect utility function should satisfy

$$\begin{aligned} 0 = \sup_{\pi,C}\Bigg\{ & \frac{\beta(1-\gamma)J}{1-1/\psi}\left[\left(\frac{C}{((1-\gamma)J)^{1/(1-\gamma)}}\right)^{1-1/\psi} - 1\right] \\ & + \big[\pi(m-r)W + rW - C\big]J_W + \frac{1}{2}\pi^2 W^2\sigma^2 J_{WW} + \kappa(\theta - r)J_r \\ & + \frac{1}{2}\delta^2 J_{rr} + \pi\rho\delta\sigma W J_{rW}\Bigg\} \end{aligned} \tag{37}$$

omitting the time subscripts t.

The first-order condition is

$$C_t = (J_W)^{-\psi}\big[(1-\gamma)J\big]^{(1-\gamma\psi)/(1-\gamma)}\beta^\psi,$$
$$\pi_t = -\frac{1}{\sigma^2 W J_{WW}}\big((m-r)J_W + \rho\delta\sigma J_{rW}\big).$$

Substituting this into (37), we get the Bellmen equation

$$0 = \frac{\beta(1-\gamma)}{1-1/\psi}J\left[\left(\frac{(J_W)^{-\psi}[(1-\gamma)J]^{(1-\gamma\psi)/(1-\gamma)}\beta^\psi}{((1-\gamma)J)^{1/(1-\gamma)}}\right)^{1-1/\psi} - 1\right]$$

$$- (J_W)^{1-\psi}\big[(1-\gamma)J\big]^{(1-\gamma\psi)/(1-\gamma)}\beta^{\psi} + \kappa(\theta - r)J_r + \frac{1}{2}\delta^2 J_{rr} + rWJ_W - \frac{1}{2}\frac{1}{\sigma^2 J_{WW}}\big((m-r)J_W + \rho\delta\sigma J_{rW}\big)^2.$$

If we guess the solution form as $J_t = I(r_t)\frac{W_t^{1-\gamma}}{1-\gamma}$ and substitute it into the Bellman equation, then we get the following ODE

$$0 = -\frac{1}{1-\psi}\beta^{\psi} I^{1+(1-\psi)/(1-\gamma)} + \frac{\psi}{1-\psi}\beta I + \frac{(m-r)^2}{2\gamma\sigma^2} I + \frac{\rho\delta(m-r)}{\gamma\sigma} I_r + \frac{\rho^2\delta^2}{2\gamma}\frac{I_r^2}{I} + \frac{\kappa(\theta-r)}{1-\gamma} I_r + \frac{\delta^2}{2(1-\gamma)} I_{rr} + rI \tag{38}$$

again, all subscripts were omitted.

5.2.1. *Exact solution when $\psi = 1$*

When $\psi \to 1$, Equation (38) can be simplified to

$$0 = \beta\log\beta I - \beta I - \frac{\beta}{1-\gamma} I\log I + \frac{(m-r)^2}{2\gamma\sigma^2} I + \frac{\rho\delta(m-r)}{\gamma\sigma} I_r + \frac{\rho^2\delta^2}{2\gamma}\frac{I_r^2}{I} + \frac{\kappa(\theta-r)}{1-\gamma} I_r + \frac{\delta^2}{2(1-\gamma)} I_{rr} + rI. \tag{39}$$

This ODE has a solution for I in the form $I = \exp(Ar_t + Br_t^2 + D)$. Plugging this guessed form into (39), we can get the equation containing A, B and D

$$0 = \beta(\log\beta - 1) - \frac{\beta}{1-\gamma}\big(Ar + Br^2 + D\big) + \frac{(m-r)^2}{2\gamma\sigma^2} + \frac{\rho\delta(m-r)(A+2Br)}{\gamma\sigma} + \frac{\rho^2\delta^2(A+2Br)^2}{2\gamma} + \frac{\kappa(\theta-r)(A+2Br)}{1-\gamma} + \frac{\delta^2(4B^2r^2 + 4ABr + A^2 + 2B)}{2(1-\gamma)} + r.$$

That is,

$$0 = \left[\frac{2\rho^2\delta^2}{\gamma} + \frac{2\delta^2}{1-\gamma}\right]B^2 - \left[\frac{2\rho\delta}{\gamma\sigma} + \frac{\beta + 2\kappa}{1-\gamma}\right]B + \frac{1}{2\gamma\sigma^2},$$

$$0 = \left[\frac{\rho^2\delta^2}{\gamma}B + \frac{2\delta^2}{1-\gamma}B - \frac{\rho\delta}{\gamma\sigma} - \frac{\beta+\kappa}{1-\gamma}\right]A + \left[\frac{2\kappa\theta}{1-\gamma} + \frac{2\rho\delta m}{\gamma\sigma}\right]B - \frac{m}{\gamma\sigma^2} + 1,$$

$$0 = -\frac{\beta}{1-\gamma}D + \left[\frac{\rho^2\delta^2}{2\gamma} + \frac{\delta^2}{2(1-\gamma)}\right]A^2 + \left[\frac{\rho\delta m}{\gamma\sigma} + \frac{\kappa\theta}{1-\gamma}\right]A + \frac{\delta^2}{1-\gamma}B + \beta(\log\beta - 1) + \frac{m^2}{2\gamma\sigma^2}.$$

Therefore, we successfully reduce the problem to a quadratic equation. Now given we have the suitable solutions for (24), the corresponding optimal asset allocation is

$$\frac{C_t}{W_t} = \beta,$$
$$\pi_t = \frac{1}{\gamma}\left(\frac{m - r_t}{\sigma^2}\right) + \frac{\rho\delta}{\gamma\sigma}[A + 2Br_t].$$

This result is intuitive. It shows that an individual is under the unit elasticity assumption indifferent to the allocation of consumption over time. Therefore, only the time preference parameter β can affect the consumption/investment ratio of an individual.

5.2.2. Approximation methods when $\psi \neq 1$

However, similar to Proposition 4, if the unit elasticity assumption does not hold, we do not have the analytic solution for (23). So we have to use the approximation methods shown above.

5.2.2.1. Perturbation method 1 (around unit elasticity). If we think the elasticity ψ is around 1, we can linearize the first term in (23)

$$I^{1+(1-\psi)/(1-\gamma)} \approx I + \frac{1-\psi}{1-\gamma} I \log I.$$

Substituting this linearized term into (38), yields the following analytically solvable equation

$$\begin{aligned} 0 = &-\frac{\beta^\psi - \psi\beta}{1-\psi} I - \frac{\beta^\psi}{1-\gamma} I \log I + \frac{(m-r)^2}{2\gamma\sigma^2} I + \frac{\rho\delta(m-r)}{\gamma\sigma} I_r \\ &+ \frac{\rho^2\delta^2}{2\gamma}\frac{I_r^2}{I} + \frac{\kappa(\theta - r)}{1-\gamma} I_r + \frac{\delta^2}{2(1-\gamma)} I_{rr} + rI. \end{aligned} \tag{40}$$

By Proposition 4, we know that this equation has an analytic solution of I in the form $I = \exp[Ar + Br^2 + D]$. Then we substitute this guessed form in and we get the equation containing A, B and D:

$$\begin{aligned} 0 = &-\frac{\beta^\psi - \psi\beta}{1-\psi} - \frac{\beta^\psi}{1-\gamma}\big(Ar + Br^2 + D\big) + \frac{(m-r)^2}{2\gamma\sigma^2} \\ &+ \frac{\rho^2\delta^2(A + 2Br)^2}{2\gamma} + \frac{\rho\delta(m-r)(A+2Br)}{\gamma\sigma} + \frac{\kappa(\theta - r)(A+2Br)}{1-\gamma} \\ &+ \frac{\delta^2(4B^2r^2 + 4ABy + A^2 + 2B)}{2(1-\gamma)} + r. \end{aligned} \tag{41}$$

Then we can solve for A, B and D:

$$
\begin{aligned}
0 &= \left[\frac{2\rho^2\delta^2}{\gamma} + \frac{2\delta^2}{1-\gamma}\right]B^2 - \left[\frac{2\rho\delta}{\gamma\sigma} + \frac{\beta^\psi + 2\kappa}{1-\gamma}\right]B + \frac{1}{2\gamma\sigma^2},\\
0 &= \left[\frac{\rho^2\delta^2}{\gamma}B + \frac{2\delta^2}{1-\gamma}B - \frac{\rho\delta}{\gamma\sigma} - \frac{\beta^\psi + \kappa}{1-\gamma}\right]A + \left[\frac{2\kappa\theta}{1-\gamma} + \frac{2\rho\delta m}{\gamma\sigma}\right]B\\
&\quad - \frac{m}{\gamma\sigma^2} + 1,\\
0 &= -\frac{\beta^\psi}{1-\gamma}D + \left[\frac{\rho^2\delta^2}{2\gamma} + \frac{\delta^2}{2(1-\gamma)}\right]A^2 + \left[\frac{\rho\delta m}{\gamma\sigma} + \frac{\kappa\theta}{1-\gamma}\right]A + \frac{\delta^2}{1-\gamma}B\\
&\quad - \frac{\beta^\psi - \psi\beta}{1-\psi} + \frac{m^2}{2\gamma\sigma^2}
\end{aligned}
$$

and the optimal consumption/investment allocation is

$$
\begin{aligned}
\frac{C_t}{W_t} &= \beta^\psi \mathrm{e}^{[Ar_t + Br_t^2 + D]\frac{1-\psi}{1-\gamma}},\\
\pi_t &= \frac{1}{\gamma}\left(\frac{m - r_t}{\sigma^2}\right) + \frac{\rho\delta}{\gamma\sigma}[A + 2Bm_t].
\end{aligned}
$$

5.2.2.2. Perturbation method 2 (around mean consumption/wealth ratio). We can further simplify Equation (38) by assuming $I = H^{-(1-\gamma)/(1-\psi)}$, then

$$
\begin{aligned}
0 &= -\beta^\psi H^{-1} + \psi\beta + \frac{(1-\psi)}{2\gamma\sigma^2}(m-r)^2 - \frac{\rho\delta(1-\gamma)}{\gamma\sigma}(m-r)\frac{H_r}{H}\\
&\quad + r(1-\psi) - \kappa(\theta - r)\frac{H_r}{H} + \frac{1}{2}\delta^2\left(\frac{1-\gamma}{1-\psi} + 1\right)\left(\frac{H_r}{H}\right)^2\\
&\quad - \frac{1}{2}\delta^2\frac{H_{rr}}{H} + \frac{\rho^2\delta^2(1-\gamma)^2}{2\gamma(1-\psi)}\left(\frac{H_r}{H}\right)^2.
\end{aligned}
\tag{42}
$$

Similar to Proposition 5, we make the following approximation to render (42) solvable

$$
\beta^\psi H^{-1} = \exp(c_t - w_t) \approx h_0 + h_1(c_t - w_t). \tag{43}
$$

Substituting (43) into (42), yields a solvable ODE with a solution in the form $H = \exp\{Ar_t + Br_t^2 + D\}$. If we substitute it in the approximated version of (42), we get

$$
\begin{aligned}
0 &= -h_0 + h_1\big(Ar + Br^2 + D - \psi\log\beta\big) + \psi\beta + \frac{1-\psi}{2\gamma\sigma^2}(m-r)^2\\
&\quad - \frac{\rho\delta(1-\gamma)}{\gamma\sigma}(m-r)(A + 2Br) + r(1-\psi) - \kappa(\theta - r)(A + 2Br)\\
&\quad + \frac{1}{2}\frac{\delta^2}{\gamma}\frac{1-\gamma}{1-\psi}\big(\gamma + \rho^2(1-\gamma)\big)(A + 2Br)^2 - \delta^2 B.
\end{aligned}
$$

Separating variables leads to a system of three algebraic equations for A, B and D, which can easily be solved in closed-form

$$
\begin{aligned}
0 &= \frac{1-\psi}{2\gamma\sigma^2} + \left(2\kappa + h_1 + \frac{2\rho\delta(1-\gamma)}{\gamma\sigma}\right)B + \frac{2\delta^2}{\gamma}\frac{1-\gamma}{1-\psi}\big(\gamma + \rho^2(1-\gamma)\big)B^2,\\
0 &= -\frac{(1-\psi)m}{\gamma\sigma^2} + (1-\psi) - 2\left[\frac{\rho\delta(1-\gamma)}{\gamma\sigma}m - \kappa\theta\right]B\\
&\quad + \left[\frac{2\delta^2}{\gamma}\frac{1-\gamma}{1-\psi}\big(\gamma + \rho^2(1-\gamma)\big)B + h_1 + \kappa + \frac{\rho\delta(1-\gamma)}{\gamma\sigma}\right]A,\\
0 &= -h_0 + h_1(D - \psi\log\beta) + \psi\beta + \frac{1-\psi}{2\gamma\sigma^2}m^2 - \frac{\rho\delta(1-\gamma)}{\gamma\sigma}mA - \kappa\theta A\\
&\quad + \frac{1}{2}\frac{\delta^2}{\gamma}\frac{1-\gamma}{1-\psi}\big(\gamma + \rho^2(1-\gamma)\big)A^2 - \delta^2 B.
\end{aligned}
$$

So the approximate indirect utility function is

$$
J(W_t, r_t) = \exp\left[-\frac{1-\gamma}{1-\psi}\big(Ar_t + Br_t^2 + D\big)\right]\frac{W_t^{1-\gamma}}{1-\gamma}.
$$

We know the optimal consumption ratio and portfolio allocation immediately:

$$
\begin{aligned}
\frac{C_t}{W_t} &= \beta^{\psi}\exp\big(-Ar_t - Br_t^2 - D\big),\\
\pi_t &= \frac{1}{\gamma}\frac{m - r_t}{\sigma^2} + \frac{\rho\delta}{\gamma\sigma}\frac{1-\gamma}{1-\psi}[A + 2rBr_t].
\end{aligned}
$$

6. Conclusions

This chapter has explored two perturbation methods for solving dynamic consumption/portfolio allocation problems under general preferences and a time-varying investment opportunity set. The restrictions that we impose are that preferences can be represented by a general recursive utility formulation and that the investment opportunity set is affine. One important result for general recursive preferences is that the Sharpe ratio is sufficient to capture the time variation in the investment opportunity set.

These fairly general restrictions prohibit the application of PDEs in order to solve the stochastic control problem. The perturbation methods allow for a quick calculation of the approximate closed-form solutions. This represents an advantage over numerical methods that require extremely time-consuming calculations with an increasing dimensionality of the problem. Both methods require a perturbation around an exact solution.

The first perturbation method is a perturbation around the exact solution when the elasticity of substitution is equal to the value of 1. We build on the exact solution that exists for the unit elasticity case in order to show that an approximate solution exists to the general problem with arbitrary elasticity of substitution.

The second perturbation method is a perturbation around the mean consumption-to-wealth ratio of the solution. We have shown that the solution to the general problem has the unconditional mean of the consumption–wealth ratio in the Bellman equation. Therefore, we substitute the unconditional mean of the consumption–wealth ratio in place of the actual consumption wealth ratio and derive a closed-form solution around which a perturbation can be taken. This perturbation reflects the difference between the realized consumption–wealth ratio and its unconditional mean and leads to an approximate analytic solution to the general consumption/portfolio allocation problem.

We prove the validity of the perturbation methods under time-varying expected returns, volatility and interest rates. We have also applied the two perturbation methods to real data under time varying volatility and the results have shown that the approximations are accurate. In addition, we show formally that the solution of both methods converge to the exact solution if the elasticity of intertemporal substitution equals one.

In our practical applications we find that the first perturbation method is advantageous in cases where a closed-form solution is important and one is only concerned with elasticities of substitution equal or close to one. However, in cases where a closed-form solution is not critical and the elasticity of substitution diverges substantially from one, the second perturbation method is the more appropriate choice.

References

Campbell, J.Y., 1993. Intertemporal asset pricing without consumption data. American Economic Review 83, 487–512.

Campbell, J.Y., Viceira, L.M., 1999. Consumption and portfolio decisions when expected returns are time varying. Quarterly Journal of Economics 114, 433–495.

Campbell, J.Y., Viceira, L.M., 2001. Who should buy long-term bonds? American Economic Review 91, 99–127.

Campbell, J.Y., Chacko, G., Rodriguez, J., Viceira, L.M., 2004. Strategic asset allocation in a continuous-time VAR model. Journal of Economic Dynamics and Control.

Chacko, G., Viceira, L.M., 2003. Spectral GMM estimation of continuous-time processes. Journal of Econometrics 116, 259–292.

Chacko, G., Viceira, L.M., 2005. Dynamic consumption and portfolio choice with stochastic volatility in incomplete markets. Review of Financial Studies 18 (4).

Dai, Q., Singleton, K.J., 2000. Specification analysis of affine term structure models. Journal of Finance 55, 1943–1978.

Duffie, D., Epstein, L.G., 1992a. Stochastic differential utility. Econometrica 60, 353–394.

Duffie, D., Epstein, L.G., 1992b. Asset pricing with stochastic differential utility. Review of Financial Studies 5, 411–436.

Duffie, D., Kan, R., 1996. A yield factor model of interest rates. Mathematical Finance 6, 379–406.

Epstein, L., Zin, S., 1989. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: A theoretical framework. Econometrica 57, 937–969.

Epstein, L., Zin, S., 1991. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: An empirical investigation. Journal of Political Economy 99, 263–286.

Fisher, M., Gilles, C., 1998. Consumption and asset prices with recursive preferences, unpublished paper, Board of Governors of the Federal Reserve System.

Giovannini, A., Weil, Ph., 1989. Risk aversion and intertemporal substitution in the capital asset pricing model, NBER Working Paper No. 2824.

Hakansson, N.H., Ziemba, W.T., 1995. Capital growth theory. In: Jarrow, R.A., Maksimovic, V., Ziemba, W.T. (Eds.), Finance Handbook. North-Holland, Amsterdam, pp. 123–144.

Kim, T.S., Omberg, E., 1996. Dynamic nonmyopic portfolio behavior. Review of Financial Studies 9, 141–161.

Liu, J., 1999. Portfolio selection in stochastic environments, unpublished paper, Stanford University.

MacLean, L., Ziemba, W.T., 1991. Growth-security profiles in capital accumulation under risk. Annals of Operations Research 31, 501–510.

MacLean, L., Ziemba, W.T., 1999. Growth versus security tradeoffs in dynamic investment analysis. Annals of Operations Research 85, 193–227.

MacLean, L., Ziemba, W.T., Li, Y., 2005. Time to wealth goals in capital accumulation. Quantitative Finance 5 (4), 343–357.

Merton, R.C., 1969. Lifetime portfolio selection under uncertainty: The continuous time case. Review of Economics and Statistics 51, 247–257.

Merton, R.C., 1971. Optimum consumption and portfolio rules in a continuous-time model. Journal of Economic Theory 3, 373–413.

Merton, R.C., 1973. An intertemporal capital asset pricing model. Econometrica 41, 867–887.

Nielsen, L., Vassalou, M., 2000. Portfolio selection with randomly time-varying moments: The role of the instantaneous capital market line, unpublished paper.

Nielsen, L., Vassalou, M., 2004. Sharpe ratios and alphas in continuous time. Journal of Financial and Quantitative Analysis 39, 103–114.

Viceira, L.M., 2001. Optimal portfolio choice for long-horizon investors with nontradable labor income. The Journal of Finance 56 (2), 433–470.

Wachter, J., 1999. Portfolio and consumption decisions under mean-reverting returns: An exact solution for complete markets, unpublished paper, Harvard University.

Chapter 9

THE KELLY CRITERION IN BLACKJACK SPORTS BETTING, AND THE STOCK MARKET*

EDWARD O. THORP

Edward O. Thorp and Associates, Newport Beach, CA 92660, USA

Contents

Abstract 386
Keywords 386
1. Introduction 387
2. Coin tossing 388
3. Optimal growth: Kelly criterion formulas for practitioners 392
   3.1. The probability of reaching a fixed goal on or before n trials 392
   3.2. The probability of ever being reduced to a fraction x of this initial bankroll 394
   3.3. The probability of being at or above a specified value at the end of a specified number of trials 395
   3.4. Continuous approximation of expected time to reach a goal 396
   3.5. Comparing fixed fraction strategies: the probability that one strategy leads another after n trials 396
4. The long run: when will the Kelly strategy "dominate"? 398
5. Blackjack 399
6. Sports betting 401
7. Wall street: the biggest game 405
   7.1. Continuous approximation 406
   7.2. The (almost) real world 409
   7.3. The case for "fractional Kelly" 411
   7.4. A remarkable formula 414
8. A case study 415
   8.1. The constraints 416
   8.2. The analysis and results 416
   8.3. The recommendation and the result 417
   8.4. The theory for a portfolio of securities 418

* Paper presented at: The 10th International Conference on Gambling and Risk Taking, Montreal, June 1997, published in: Finding the Edge: Mathematical Analysis of Casino Games, edited by O. Vancura, J.A. Cornelius, W.R. Eadington, 2000. Corrections added April 20, 2005.

Handbook of Asset and Liability Management, Volume 1
Edited by S.A. Zenios and W.T. Ziemba
© 2006 Published by Elsevier B.V.
DOI: 10.1016/S1872-0978(06)01009-X

9. My experience with the Kelly approach 419
10. Conclusion 420
Acknowledgements 420
Appendix A. Integrals for deriving moments of $\boldsymbol{E}_\infty$ 420
Appendix B. Derivation of formula (3.1) 421
Appendix C. Expected time to reach goal 423
References 428

Abstract

The central problem for gamblers is to find positive expectation bets. But the gambler also needs to know how to manage his money, i.e., how much to bet. In the stock market (more inclusively, the securities markets) the problem is similar but more complex. The gambler, who is now an "investor", looks for "excess risk adjusted return". In both these settings, we explore the use of the Kelly criterion, which is to maximize the expected value of the logarithm of wealth ("maximize expected logarithmic utility"). The criterion is known to economists and financial theorists by names such as the "geometric mean maximizing portfolio strategy", maximizing logarithmic utility, the growth-optimal strategy, the capital growth criterion, etc. The author initiated the practical application of the Kelly criterion by using it for card counting in blackjack. We will present some useful formulas and methods to answer various natural questions about it that arise in blackjack and other gambling games. Then we illustrate its recent use in a successful casino sports betting system. Finally, we discuss its application to the securities markets where it has helped the author to make a thirty year total of 80 billion dollars worth of "bets".

Keywords

Kelly criterion, betting, long run investing, portfolio allocation, logarithmic utility, capital growth

JEL classification: C61, D81, G1

1. Introduction

The fundamental problem in gambling is to find positive expectation betting opportunities. The analogous problem in investing is to find investments with excess risk-adjusted expected rates of return. Once these favorable opportunities have been identified, the gambler or investor must decide how much of his capital to bet. This is the problem which we consider here. It has been of interest at least since the eighteenth century discussion of the St. Petersburg Paradox (Feller, 1966) by Daniel Bernoulli.

One approach is to choose a goal, such as to minimize the probability of total loss within a specified number of trials, N. Another example would be to maximize the probability of reaching a fixed goal on or before N trials (Browne, 1996).

A different approach, much studied by economists and others, is to value money using a utility function. These are typically defined for all non-negative real numbers, have extended real number values, and are non-decreasing (more money is at least as good as less money). Some examples are $U(x) = x^a$, $0 \leqslant a < \infty$, and $U(x) = \log x$, where log means $\log_e$, and $\log 0 = -\infty$. Once a utility function is specified, the object is to maximize the expected value of the utility of wealth.

Daniel Bernoulli used the utility function $\log x$ to "solve" the St. Petersburg Paradox. (But his solution does not eliminate the paradox because every utility function which is unbounded above, including log, has a modified version of the St. Petersburg Paradox.) The utility function $\log x$ was revisited by Kelly (1956) where he showed that it had some remarkable properties. These were elaborated and generalized in an important paper by Breiman (1961). Markowitz (1959) illustrates the application to securities. For a discussion of the Kelly criterion (the "geometric mean criterion") from a finance point of view, see McEnally (1986). He also includes additional history and references.

I was introduced to the Kelly paper by Claude Shannon at M.I.T. in 1960, shortly after I had created the mathematical theory of card counting at casino blackjack. Kelly's criterion was a bet on each trial so as to maximize $E \log X$, the expected value of the logarithm of the (random variable) capital X. I used it in actual play and introduced it to the gambling community in the first edition of Beat the Dealer (Thorp, 1962). If all blackjack bets paid even money, had positive expectation and were independent, the resulting Kelly betting recipe when playing one hand at a time would be extremely simple: bet a fraction of your current capital equal to your expectation. This is modified somewhat in practice (generally down) to allow for having to make some negative expectation "waiting bets", for the higher variance due to the occurrence of payoffs greater than one to one, and when more than one hand is played at a time.

Here are the properties that made the Kelly criterion so appealing. For ease of understanding, we illustrate using the simplest case, coin tossing, but the concepts and conclusions generalize greatly.

2. Coin tossing

Imagine that we are faced with an infinitely wealthy opponent who will wager even money bets made on repeated independent trials of a biased coin. Further, suppose that on each trial our win probability is $p > 1/2$ and the probability of losing is $q = 1 - p$. Our initial capital is X_0. Suppose we choose the goal of maximizing the expected value $E(X_n)$ after n trials. How much should we bet, B_k, on the kth trial? Letting $T_k = 1$ if the kth trial is a win and $T_k = -1$ if it is a loss, then $X_k = X_{k-1} + T_k B_k$ for $k = 1, 2, 3, \ldots$, and $X_n = X_0 + \sum_{k=1}^{n} T_k B_k$. Then

$$E(X_n) = X_0 + \sum_{k=1}^{n} E(B_k T_k) = X_0 + \sum_{k=1}^{n} (p - q) E(B_k).$$

Since the game has a positive expectation, i.e., $p - q > 0$ in this even payoff situation, then in order to maximize $E(X_n)$ we would want to maximize $E(B_k)$ at each trial. Thus, to maximize expected gain we should bet *all of our resources* at each trial. Thus $B_1 = X_0$ and if we win the first bet, $B_2 = 2X_0$, etc. However, the probability of ruin is given by $1 - p^n$ and with $p < 1$, $\lim_{n\to\infty}[1 - p^n] = 1$ so ruin is almost sure. Thus the "bold" criterion of betting to maximize expected gain is usually undesirable.

Likewise, if we play to minimize the probability of eventual ruin (i.e., "ruin" occurs if $X_k = 0$ on the kth outcome) the well-known gambler's ruin formula in Feller (1966) shows that we minimize ruin by making a *minimum* bet on each trial, but this unfortunately also minimizes the expected gain. Thus "timid" betting is also unattractive.

This suggests an intermediate strategy which is somewhere between maximizing $E(X_n)$ (and assuring ruin) and minimizing the probability of ruin (and minimizing $E(X_n)$). An asymptotically optimal strategy was first proposed by Kelly (1956).

In the coin-tossing game just described, since the probabilities and payoffs for each bet are the same, it seems plausible that an "optimal" strategy will involve always wagering the same fraction f of your bankroll. To make this possible we shall assume from here on that capital is infinitely divisible. This assumption usually does not matter much in the interesting practical applications.

If we bet according to $B_i = f X_{i-1}$, where $0 \leqslant f \leqslant 1$, this is sometimes called "fixed fraction" betting. Where S and F are the number of successes and failures, respectively, in n trials, then our capital after n trials is $X_n = X_0(1+f)^S(1-f)^F$, where $S+F = n$. With f in the interval $0 < f < 1$, $\Pr(X_n = 0) = 0$. Thus "ruin" in the technical sense of the gambler's ruin problem cannot occur. "Ruin" shall henceforth be reinterpreted to mean that for arbitrarily small positive ε, $\lim_{n\to\infty}[\Pr(X_n \leqslant \varepsilon)] = 1$. Even in this sense, as we shall see, ruin *can* occur under certain circumstances.

We note that since

$$e^{n \log \left[\frac{X_n}{X_0}\right]^{1/n}} = \frac{X_n}{X_0},$$

the quantity

$$G_n(f) = \log\left[\frac{X_n}{X_0}\right]^{1/n} = \frac{S}{n}\log(1+f) + \frac{F}{n}\log(1-f)$$

measures the exponential rate of increase per trial. Kelly chose to maximize the expected value of the growth rate coefficient, $g(f)$, where

$$\begin{aligned} g(f) &= E\left\{\log\left[\frac{X_n}{X_0}\right]^{1/n}\right\} = E\left\{\frac{S}{n}\log(1+f) + \frac{F}{n}\log(1-f)\right\} \\ &= p\log(1+f) + q\log(1-f). \end{aligned}$$

Note that $g(f) = (1/n)E(\log X_n) - (1/n)\log X_0$ so for n fixed, maximizing $g(f)$ is the same as maximizing $E\log X_n$. We usually will talk about maximizing $g(f)$ in the discussion below. Note that

$$g'(f) = \frac{p}{1+f} - \frac{q}{1-f} = \frac{p-q-f}{(1+f)(1-f)} = 0$$

when $f = f^* = p - q$.

Now

$$g''(f) = -p/(1+f)^2 - q/(1-f)^2 < 0$$

so that $g'(f)$ is monotone strictly decreasing on $[0, 1)$. Also $g'(0) = p - q > 0$ and $\lim_{f\to 1^-} g'(f) = -\infty$. Therefore by the continuity of $g'(f)$, $g(f)$ has a unique maximum at $f = f^*$, where $g(f^*) = p\log p + q\log q + \log 2 > 0$. Moreover, $g(0) = 0$ and $\lim_{f\to q^-} g(f) = -\infty$ so there is a unique number $f_c > 0$, where $0 < f^* < f_c < 1$, such that $g(f_c) = 0$. The nature of the function $g(f)$ is now apparent and a graph of $g(f)$ versus f appears as shown in Figure 1.

The following theorem recounts the important advantages of maximizing $g(f)$. The details are omitted here but proofs of (i)–(iii), and (vi) for the simple binomial case can be found in Thorp (1969); more general proofs of these and of (iv) and (v) are in Breiman (1961).

Theorem 1. (i) *If* $g(f) > 0$, *then* $\lim_{n\to\infty} X_n = \infty$ *almost surely, i.e., for each* M, $\Pr[\liminf_{n\to\infty} X_n > M] = 1$;

(ii) *If* $g(f) < 0$, *then* $\lim_{n\to\infty} X_n = 0$ *almost surely*; *i.e., for each* $\varepsilon > 0$, $\Pr[\limsup_{n\to\infty} X_n < \varepsilon] = 1$;

(iii) *If* $g(f) = 0$, *then* $\limsup_{n\to\infty} X_n = \infty$ *a.s. and* $\liminf_{n\to\infty} X_n = 0$ *a.s.*

(iv) *Given a strategy* Φ^* *which maximizes* $E\log X_n$ *and any other "essentially different" strategy* Φ (*not necessarily a fixed fractional betting strategy*), *then* $\lim_{n\to\infty} X_n(\Phi^*)/X_n(\Phi) = \infty$ *a.s.*

(v) *The expected time for the current capital* X_n *to reach any fixed preassigned goal* C *is, asymptotically, least with a strategy which maximizes* $E\log X_n$.

(vi) *Suppose the return on one unit bet on the* ith *trial is the binomial random variable* U_i; *further, suppose that the probability of success is* p_i, *where* $1/2 < p_i < 1$.

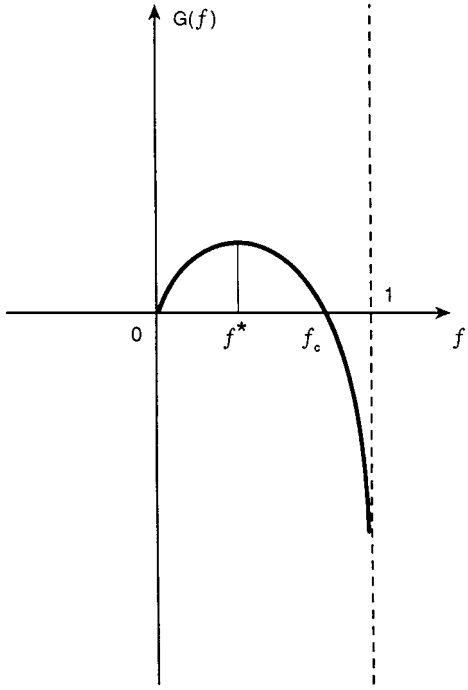


Fig. 1.

Then $E \log X_n$ *is maximized by choosing on each trial the fraction* $f_i^* = p_i - q_i$ *which maximizes* $E \log(1 + f_i U_i)$.

Part (i) shows that, except for a finite number of terms, the player's fortune X_n will exceed any fixed bound M when f is chosen in the interval $(0, f_c)$. But, if $f > f_c$, part (ii) shows that ruin is almost sure. Part (iii) demonstrates that if $f = f_c$, X_n will (almost surely) oscillate randomly between 0 and $+\infty$. Thus, one author's statement that $X_n \to X_0$ as $n \to \infty$, when $f = f_e$, is clearly contradicted. Parts (iv) and (v) show that the Kelly strategy of maximizing $E \log X_n$ is asymptotically optimal by two important criteria. An "essentially different" strategy is one such that the difference $E \ln X_n^* - E \ln X_n$ between the Kelly strategy and the other strategy grows faster than the standard deviation of $\ln X_n^* - \ln X_n$, ensuring $P(\ln X_n^* - \ln X_n > 0) \to 1$. Part (vi) establishes the validity of utilizing the Kelly method of choosing f_i^* on each trial (even if the probabilities change from one trial to the next) in order to maximize $E \log X_n$.

Example 2.1. Player A plays against an infinitely wealthy adversary. Player A wins even money on successive independent flips of a biased coin with a win probability of $p = .53$ (no ties). Player A has an initial capital of X_0 and *capital is infinitely divisible*.

Applying Theorem 1(vi), $f^* = p - q = .53 - .47 = .06$. Thus 6% of current capital should be wagered on each play in order to cause X_n to grow at the fastest rate possible consistent with zero probability of ever going broke. If Player A continually bets a fraction smaller than 6%, X_n will also grow to infinity but the rate will be slower.

If Player A repeatedly bets a fraction larger than 6%, up to the value f_c, the same thing applies. Solving the equation $g(f) = .53\log(1+f) + .47\log(1-f) = 0$ numerically on a computer yields $f_c = .11973^-$. So, if the fraction wagered is more than about 12%, then even though Player A may temporarily experience the pleasure of a faster win rate, eventual downward fluctuations will inexorably drive the values of X_n toward zero. Calculation yields a growth coefficient of $g(f^*) = f(.06) = .001801$ so that after n successive bets the log of Player A's average bankroll will tend to $.001801n$ times as much money as he started with. Setting $.001801n = \log 2$ gives an expected time of about $n = 385$ to double the bankroll.

The Kelly criterion can easily be extended to uneven payoff games. Suppose Player A wins b units for every unit wager. Further, suppose that on each trial the win probability is $p > 0$ and $pb - q > 0$ so the game is advantageous to Player A. Methods similar to those already described can be used to maximize

$$g(f) = E\log(X_n/X_0) = p\log(1+bf) + q\log(1-f).$$

Arguments using calculus yield $f^* = (bp-q)/b$, the optimal fraction of current capital which should be wagered on each play in order to maximize the growth coefficient $g(f)$.

This formula for f^* appeared in Thorp (1984) and was the subject of an April 1997 discussion on the Internet at Stanford Wong's website, http://bj21.com (miscellaneous free pages section). One claim was that one can only lose the amount bet so there was no reason to consider the (simple) generalization of this formula to the situation where a unit wager wins b with probability $p > 0$ and loses a with probability q. Then if the expectation $m \equiv bp - aq > 0$, $f^* > 0$ and $f^* = m/ab$. The generalization does stand up to the objection. One can buy on credit in the financial markets and lose much more than the amount bet. Consider buying commodity futures or selling short a security (where the loss is potentially unlimited). See, e.g., Thorp and Kassouf (1967) for an account of the E.L. Bruce short squeeze.

For purists who insist that these payoffs are not binary, consider selling short a binary digital option. These options are described in Browne (1996).

A criticism sometimes applied to the Kelly strategy is that capital is not, in fact, infinitely divisible. In the real world, bets are multiples of a minimum unit, such as \$1 or \$.01 (penny "slots"). In the securities markets, with computerized records, the minimum unit can be as small as desired. With a minimum allowed bet, "ruin" in the standard sense is always possible. It is not difficult to show, however (see Thorp and Walden, 1966) that if the minimum bet allowed is small relative to the gambler's initial capital, then the probability of ruin in the standard sense is "negligible" and also that the theory herein described is a useful approximation. This section follows Rotando and Thorp (1992).

3. Optimal growth: Kelly criterion formulas for practitioners

Since the Kelly criterion asymptotically maximizes the expected growth rate of wealth, it is often called the optimal growth strategy. It is interesting to compare it with the other fixed fraction strategies. I will present some results that I have found useful in practice. My object is to do so in a way that is simple and easily understood. These results have come mostly from sitting and thinking about "interesting questions". I have not made a thorough literature search but I know that some of these results have been previously published and in greater mathematical generality. See, e.g., Browne (1996, 1997) and the references therein.

3.1. The probability of reaching a fixed goal on or before n trials

We first assume coin tossing. We begin by noting a related result for standard Brownian motion. Howard Tucker showed me this in 1974 and it is probably the most useful single fact I know for dealing with diverse problems in gambling and in the theory of financial derivatives.

For standard Brownian motion $X(t)$, we have

$$P\big(\sup\big[X(t) - (at+b)\big] \geqslant 0,\ 0 \leqslant t \leqslant T\big) = N(-\alpha-\beta) + \mathrm{e}^{-2ab}N(\alpha-\beta) \tag{3.1}$$

where $\alpha = a\sqrt{T}$ and $\beta = b/\sqrt{T}$. See Figure 2. See Appendix B for Tucker's derivation of (3.1).

In our application $a < 0, b > 0$ so we expect $\lim_{T\to\infty} P(X(t) \geqslant at+b,\ 0 \leqslant t \leqslant T) = 1$.

Let f be the fraction bet. Assume independent identically distributed (i.d.d.) trials Y_i, $i = 1, \ldots, n$, with $P(Y_i = 1) = p > 1/2$, $P(Y_i = -1) = q < 1/2$; also assume $p < 1$ to avoid the trivial case $p = 1$.

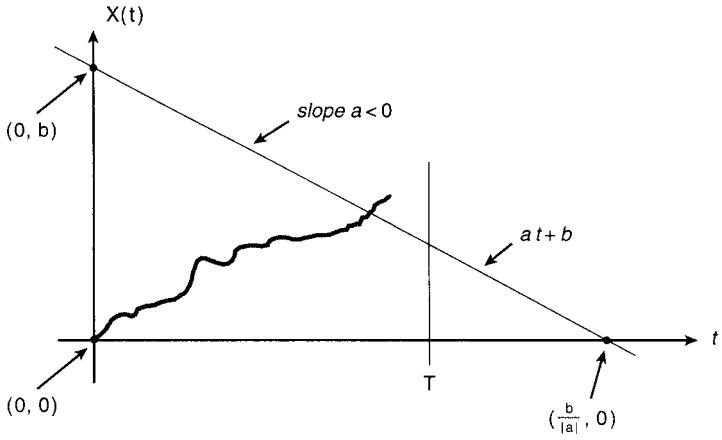


Fig. 2.

Bet a fixed fraction f, $0 < f < 1$, at each trial. Let V_k be the value of the gambler or investor's bankroll after k trials, with initial value V_0. Choose initial stake $V_0 = 1$ (without loss of generality); number of trials n; goal $C > 1$.

What is the probability that $V_k \geqslant C$ for some k, $1 \leqslant k \leqslant n$? This is the same as the probability that $\log V_k \geqslant \log C$ for some k, $1 \leqslant k \leqslant n$. Letting $\ln = \log_e$ we have:

$$V_k = \prod_{i-1}^{k}(1 + Y_i f) \quad \text{and}$$

$$\ln V_k = \sum_{i=1}^{k} \ln(1 + Y_i f),$$

$$E \ln V_k = \sum_{i=1}^{k} E \ln(1 + Y_i f),$$

$$\operatorname{Var}(\ln V_k) = \sum_{i=1}^{k} \operatorname{Var}\big(\ln(1 + Y_i f)\big),$$

$$E \ln(1 + Y_i f) = p \ln(1 + f) + q \ln(1 - f) \equiv m \equiv g(f),$$

$$\begin{aligned}
\operatorname{Var}\big[\ln(1 + Y_i f)\big] &= p\big[\ln(1 + f)\big]^2 + q\big[\ln(1 - f)\big]^2 - m^2 \\
&= \big(p - p^2\big)\big[\ln(1 + f)\big]^2 + \big(q - q^2\big)\big[\ln(1 - f)\big]^2 \\
&\quad - 2pq \ln(1 + f) \ln(1 - f) \\
&= pq\big\{\big[\ln(1 + f)\big]^2 - 2 \ln(1 + f) \ln(1 - f) + \big[\ln(1 - f)\big]^2\big\} \\
&= pq\big\{\ln\big[(1 + f)/(1 - f)\big]\big\}^2 \equiv s^2.
\end{aligned}$$

Drift in n trials: mn.
Variance in n trials: $s^2 n$.

$$\ln V_k \geqslant \ln C, \ 1 \leqslant k \leqslant n, \quad \text{iff}$$

$$\sum_{i=1}^{k} \ln(1 + Y_i f) \geqslant \ln C, \ 1 \leqslant k \leqslant n, \quad \text{iff}$$

$$S_k \equiv \sum_{i=1}^{k} \big[\ln(1 + Y_i f) - m\big] \geqslant \ln C - mk, \quad 1 \leqslant k \leqslant n,$$

$$E(S_k) = 0, \quad \operatorname{Var}(S_k) = s^2 k.$$

We want $\operatorname{Prob}(S_k \geqslant \ln C - mk, \ 1 \leqslant k \leqslant n)$.

Now we use our Brownian motion formula to approximate S_n by $\operatorname{Prob}(X(t) \geqslant \ln C - mt/s^2, \ 1 \leqslant t \leqslant s^2 n)$ where each term of S_n is approximated by an $X(t)$, drift 0 and

variance s^2 ($0 \leqslant t \leqslant s^2$, $s^2 \leqslant t \leqslant 2s^2$, $\ldots$, $(n-1)s^2 \leqslant t \leqslant ns^2$). Note: the approximation is only "good" for "large" n.

Then in the original formula (3.1):

$$T = s^2 n,$$
$$b = \ln C,$$
$$a = -m/s^2,$$
$$\alpha = a\sqrt{T} = -m\sqrt{n}/s,$$
$$\beta = b/\sqrt{T} = \ln C/s\sqrt{n}.$$

Example 3.1.

$$C = 2,$$
$$n = 10^4,$$
$$p = .51,$$
$$q = .49,$$
$$f = .0117,$$
$$m = .000165561,$$
$$s^2 = .000136848.$$

Then

$$P(\cdot) = .9142.$$

Example 3.2. Repeat with

$$f = .02,$$

then

$$m = .000200013, \quad s^2 = .000399947 \quad \text{and} \quad P(\cdot) = .9214.$$

3.2. The probability of ever being reduced to a fraction x of this initial bankroll

This is a question that is of great concern to gamblers and investors. It is readily answered, approximately, by our previous methods.

Using the notation of the previous section, we want $P(V_k \leqslant x \text{ for some } k,\ 1 \leqslant k \leqslant \infty)$. Similar methods yield the (much simpler) continuous approximation formula:

$$\text{Prob}(\cdot) = \mathrm{e}^{2ab} \quad \text{where } a = -m/s^2 \text{ and } b = -\ln x$$

which can be rewritten as

$$\text{Prob}(\cdot) = x^{\wedge}(2m/s^2) \quad \text{where } \wedge \text{ means exponentiation.} \tag{3.2}$$

Example 3.3.

$$p = .51, \quad f = f^* = .02,$$
$$2m/s = 1.0002,$$
$$\text{Prob}(\cdot) \doteq x.$$

We will see in Section 7 that for the limiting continuous approximation and the Kelly optimal fraction f^*, $P(V_k(f^*) \leqslant x \text{ for some } k \geqslant 1) = x$.

My experience has been that most cautious gamblers or investors who use Kelly find the frequency of substantial bankroll reduction to be uncomfortably large. We can see why now. To reduce this, they tend to prefer somewhat less than the full betting fraction f^*. This also offers a margin of safety in case the betting situations are less favorable than believed. The penalty in reduced growth rate is not severe for moderate underbetting. We discuss this further in Section 7.

3.3. The probability of being at or above a specified value at the end of a specified number of trials

Hecht (1995) suggested setting this probability as the goal and used a computerized search method to determine optimal (by this criterion) fixed fractions for $p - q = .02$ and various c, n and specified success probabilities.

This is a much easier problem than the similar sounding in Section 3.1. We have for the probability that $X(T)$ at the end exceeds the goal:

$$P\big(X(T) \geqslant aT + b\big) = \frac{1}{\sqrt{2\pi T}} \int_{aT+b}^{\infty} \exp\{-x^2/2T\}\,\mathrm{d}x$$
$$= \frac{1}{\sqrt{2\pi T}} \int_{aT^{1/2}+bT^{-1/2}}^{\infty} \exp\{-u^2/2\}\,\mathrm{d}u$$

where $u = x/\sqrt{T}$ so $x = aT + b$ gives $u\sqrt{T} = aT + b$ and $U = aT^{1/2} + bT^{-1/2}$. The integral equals

$$1 - N\big(aT^{1/2} + bT^{-1/2}\big) = N\big(-\big(aT^{1/2} + bT^{-1/2}\big)\big)$$
$$= 1 - N(\alpha + \beta) = N(-\alpha - \beta). \tag{3.3}$$

For example (3.1) $f = .0117$ and $P = .7947$. For example (3.2) $P = .7433$. Example (3.1) is for the Hecht optimal fraction and example (3.2) is for the Kelly optimal fraction. Note the difference in P values.

Our numerical results are consistent with Hecht's simulations in the instances we have checked.

Browne (1996) has given an elegant continuous approximation solution to the problem: What is the strategy which maximizes the probability of reaching a fixed goal C on or before a specified time n and what is the corresponding probability of success? Note

that the optimal strategy will in general involve betting varying fractions, depending on the time remaining and the distance to the goal.

As an extreme example, just to make the point, suppose $n = 1$ and $C = 2$. If $X_0 < 1$ then no strategy works and the probability of success is 0. But if $1 \leqslant X_0 < 2$ one should bet at least $2 - X_0$, thus any fraction $f \geqslant (2 - X_0)/X_0$, for a success probability of p. Another extreme example: $n = 10$, $C = 2^{10} = 1024$, $X_0 = 1$. Then the only strategy which can succeed is to bet $f = 1$ on every trial. The probability of success is p^{10} for this strategy and 0 for all others (if $p < 1$), including Kelly.

3.4. Continuous approximation of expected time to reach a goal

According to Theorem 1(v), the optimal growth strategy asymptotically minimizes the expected time to reach a goal. Here is what this means. Suppose for goal C that $m(C)$ is the greatest lower bound over all strategies for the expected time to reach C. Suppose $t^*(C)$ is the expected time using the Kelly strategy. Then $\lim_{C\to\infty}(t^*(c)/m(c)) = 1$.

The continuous approximation to the expected number of trials to reach the goal $C > X_0 = 1$ is

$$n(C, f) = (\ln C)/g(f)$$

where f is any fixed fraction strategy. Appendix C has the derivation. Now $g(f)$ has a unique maximum at $g(f^*)$ so $n(C, f)$ has a unique minimum at $f = f^*$. Moreover, we can see how much longer it takes, on average, to reach C if one deviates from f^*.

3.5. Comparing fixed fraction strategies: the probability that one strategy leads another after n trials

Theorem 1(iv) says that wealth using the Kelly strategy will tend, in the long run, to an infinitely large multiple of wealth using any "essentially different" strategy. It can be shown that any fixed $f \neq f^*$ is an "essentially different" strategy. This leads to the question of how fast the Kelly strategy gets ahead of another fixed fraction strategy, and more generally, how fast one fixed fraction strategy gets ahead of (or behind) another.

If W_n is the number of wins in n trials and $n - W_n$ is the number of losses,

$$G(f) = (W_n/n)\ln(l + f) + (1 - W_n/n)\ln(1 - f)$$

is the actual (random variable) growth coefficient.

As we saw, its expectation is

$$g(f) = E\big(G(f)\big) = p\log(1 + f) + q\log(1 - f) \tag{3.4}$$

and the variance of $G(f)$ is

$$\operatorname{Var} G(f) = \big((pq)/n\big)\big\{\ln\big((1 + f)/(1 - f)\big)\big\}^2 \tag{3.5}$$

and it follows that $G(f)$, which has the form $G(f) = a(\sum T_k)/n + b$, is approximately normally distributed with mean $g(f)$ and variance $\operatorname{Var} G(f)$. This enables us to give

the distribution of X_n and once again answer the question of Section 3.3. We illustrate this with an example.

Example 3.4.

$p = .51, \quad q = .49, \quad f^* = .02, \quad N = 10{,}000$ and

$s =$ standard deviation of $G(f)$

g/s	f	g	s	$\Pr(G(f) \leqslant 0)$
1.5	.01	.000150004	.0001	.067
1.0	.02	.000200013	.0002	.159
.5	.03	.000149977	.0003	.309

Continuing, we find the distribution of $G(f_2) - G(f_1)$. We consider two cases.

Case 1. The *same* game

Here we assume both players are betting on the same trials, e.g., betting on the same coin tosses, or on the same series of hands at blackjack, or on the same games with the same odds at the same sports book. In the stock market, both players could invest in the same "security" at the same time, e.g., a no-load S&P 500 index mutual fund.

We find

$$E\big(G(f_2) - G(f_1)\big) = p\log\big((1+f_2)/(1+f_1)\big) + q\log\big((1-f_2)/(1-f_1)\big)$$

and

$$\operatorname{Var}\big(G(f_2) - G(f_1)\big) = (pq/n)\left\{\log\left[\left(\frac{1+f_2}{1-f_2}\right)\left(\frac{1-f_1}{1+f_1}\right)\right]\right\}^2$$

where $G(f_2) - G(f_1)$ is approximately normally distributed with this mean and variance.

Case 2. Identically distributed independent games

This corresponds to betting on two different series of tosses with the same coin. $E(G(f_2) - G(f_1))$ is as before. But now $\operatorname{Var}(G(f_2) - G(f_1)) = \operatorname{Var}(G(f_2)) + \operatorname{Var}(G(f_1))$ because $G(f_2)$ and $G(f_1)$ are now independent. Thus

$$\operatorname{Var}\big(G(f_2) - G(f_1)\big) = (pq/n)\left\{\left[\log\left(\frac{1+f_2}{1-f_2}\right)\right]^2 + \left[\log\left(\frac{1+f_1}{1-f_1}\right)^2\right]\right\}.$$

Let

$$a = \log\left(\frac{1+f_1}{1-f_1}\right), \qquad b = \log\left(\frac{1+f_2}{1-f_2}\right).$$

Then in Case 1, $V_1 = (pq/n)(a-b)^2$ and in Case 2, $V_2 = (pq/n)(a^2+b^2)$ and since $a, b > 0$, $V_1 < V_2$ as expected. We can now compare the Kelly strategy with other

fixed fractions to determine the probability that Kelly leads after n trials. Note that this probability is always greater than $1/2$ (within the accuracy limits of the continuous approximation, which is the approximation of the binomial distribution by the normal, with its well known and thoroughly studied properties) because $g(f^*) - g(f) > 0$ where $f^* = p - q$ and $f \neq f^*$ is some alternative. This can fail to be true for small n, where the approximation is poor. As an extreme example to make the point, if $n = 1$, any $f > f^*$ beats Kelly with probability $p > 1/2$. If instead $n = 2$, $f > f^*$ wins with probability p^2 and $p^2 > 1/2$ if $p > 1/\sqrt{2} \doteq .7071$. Also, $f < f^*$ wins with probability $1 - p^2$ and $1 - p^2 > 1/2$ if $p^2 < 1/2$, i.e., $p < 1/\sqrt{2} = .7071$. So when $n = 2$, Kelly always loses more than half the time to some other f unless $p = 1/\sqrt{2}$.

We now have the formulas we need to explore many practical applications of the Kelly criterion.

4. The long run: when will the Kelly strategy "dominate"?

The late John Leib wrote several articles for Blackjack Forum which were critical of the Kelly criterion. He was much bemused by "the long run". What is it and when, if ever, does it happen?

We begin with an example.

Example 4.1.

$$p = .51, \quad n = 10{,}000,$$

V_i and s_i, $i = 1, 2$, are the variance and standard deviation, respectively, for Section 3.5 Cases 1 and 2, and $R = V_2/V_1 = (a^2 + b^2)/(a - b)^2$ so $s_2 = s_1\sqrt{R}$. Table 1 summarizes some results. We can also approximate $\sqrt{R}$ with a power series estimate using only the first term of a and of b: $a \doteq 2f_1$, $b \doteq 2f_2$ so $\sqrt{R} \doteq \sqrt{f_1^2 + f_2^2}/|f_1 - f_2|$. The approximate results, which agree extremely well, are 2.236, 3.606 and 1.581, respectively.

The first two rows show how nearly symmetric the behavior is on each side of the optimal $f^* = .02$. The column $(g_2 - g_1)/s_1$ shows us that $f^* = .02$ only has a .5 standard deviation advantage over its neighbors $f = .01$ and $f = .03$ after $n = 10{,}000$

Table 1
Comparing strategies

f_1	f_2	$g_2 - g_1$	s_1	$(g_2 - g_1)/s_1$	$\sqrt{R}$
.01	.02	.00005001	.00010000	.50	2.236
.03	.02	.00005004	.00010004	.50	3.604
.03	.01	.00000003	.00020005	.00013346	1.581

Table 2
The long run: $(g_2 - g_1)/s$ after n trials

f_1	f_2	$n = 10^4$	$n = 4 \times 10^4$	$n = 16 \times 10^4$	$n = 10^6$
.01	.02	.5	1.0	2.0	5.0
.03	.02	.5	1.0	2.0	5.0
.03	.01	.000133	.000267	.000534	.001335

trials. Since this advantage is proportional to $\sqrt{n}$, the column $(g_2 - g_1)/s_1$ from Table 1 gives the results of Table 2.

The factor $\sqrt{R}$ in Table 1 shows how much more slowly f_2 dominates f_1 in Case 2 versus Case 1. The ratio $(g_2 - g_1)/s_2$ is $\sqrt{R}$ times as large so the same level of dominance takes R times as long. When the real world comparisons of strategies for practical reasons often use Case 2 comparisons rather than the more appropriate Case 1 comparisons, the dominance of f^* is further obscured. An example is players with different betting fractions at blackjack. Case 1 corresponds to both betting on the same sequence of hands. Case 2 corresponds to them playing at different tables (not the same table, because Case 2 assumes independence). (Because of the positive correlation between payoffs on hands played at the same table, this is intermediate between Cases 1 and 2.)

It is important to understand that "the long run", i.e., the time it takes for f^* to dominate a specified neighbor by a specified probability, can vary without limit. Each application requires a separate analysis. In cases such as Example 4.1, where dominance is "slow", one might argue that using f^* is not important. As an argument against this, consider two coin-tossing games. In game 1 your edge is 1.0%. In game 2 your edge is 1.1%. With one unit bets, after n trials the difference in expected gain is $E_2 - E_1 = .001n$ with standard deviation s of about $\sqrt{2n}$ hence $(E_2 - E_1)/s \doteq .001\sqrt{n}/\sqrt{2}$ which is 1 when $n = 2 \times 10^6$. So it takes two million trials to have an 84% chance of the game 2 results being better than the game 1 results. Does that mean it's unimportant to select the higher expectation game?

5. Blackjack

For a general discussion of blackjack, see Thorp (1962, 1966), Wong (1994) and Griffin (1979). The Kelly criterion was introduced for blackjack by Thorp (1962). The analysis is more complicated than that of coin tossing because the payoffs are not simply one to one. In particular the variance is generally more than 1 and the Kelly fraction tends to be less than for coin tossing with the same expectation. Moreover, the distribution of various payoffs depends on the player advantage. For instance the frequency of pair splitting, doubling down, and blackjacks all vary as the advantage changes. By binning the probability of payoff types according to ex ante expectation, and solving the Kelly equations on a computer, a strategy can be found which is as close to optimal as desired.

There are some conceptual subtleties which are noteworthy. To illustrate them we'll simplify to the coin toss model.

At each trial, we have with probability .5 a "favorable situation" with gain or loss X per unit bet such that $P(X = 1) = .51$, $P(X = -1) = .49$ and with probability .5 an unfavorable situation with gain or loss Y per unit bet such that $P(Y = 1) = .49$ and $P(Y = -1) = .51$. We know before we bet whether X or Y applies.

Suppose the player must make small "waiting" bets on the unfavorable situations in order to be able to exploit the favorable situations. On these he will place "large" bets. We consider two cases.

Case 1. Bet f_0 on unfavorable situations and find the optimal f^* for favorable situations. We have

$$g(f) = .5\big(.51\log(1+f) + .49\log(1-f)\big) + .5\big(.49\log(1+f_0) + .51\log(1-f_0)\big). \tag{5.1}$$

Since the second expression in (5.1) is constant, f maximizes $g(f)$ if it maximizes the first expression, so $f^* = p - q = .02$, as usual. It is easy to verify that when there is a spectrum of favorable situations the same recipe, $f_i^* = p_i - q_i$ for the ith situation, holds. Again, in actual blackjack f_i^* would be adjusted down somewhat for the greater variance. With an additional constraint such as $f_i \leqslant kf_0$, where k is typically some integral multiple of f_0, representing the betting spread adopted by a prudent player, then the solution is just $f_i \leqslant \min(f_i^*, kf_0)$.

Curiously, a seemingly similar formulation of the betting problem leads to rather different results.

Case 2. Bet f in favorable situations and af in unfavorable situations, $0 \leqslant a \leqslant 1$.

Now the bet sizes in the two situations are linked and both the analysis and results are more complex. We have a Kelly growth rate of

$$g(f) = .5\big(.51\log(1+f) + .49\log(1-f)\big) + .5\big(.49\log(1+af) + .51\log(1-af)\big). \tag{5.2}$$

If we choose $a = 0$ (no bet in unfavorable situations) then the maximum value for $g(f)$ is at $f^* = .02$, the usual Kelly fraction.

If we make "waiting bets", corresponding to some value of $a > 0$, this will shift the value of f^* down, perhaps even to 0. The expected gain divided by the expected bet is $.02(1-a)/(1+a)$, $a \geqslant 0$. If $a = 0$ we get .02, as expected. If $a = 1$, we get 0, as expected: this is a fair game and the Kelly fraction is $f^* = 0$. As a increases from 0 to 1 the (optimal) Kelly fraction f^* decreases from .02 to 0. Thus the Kelly fraction for favorable situations is less *in this case* when bets on unfavorable situations reduce the overall advantage of the game.

Arnold Snyder called to my attention the fact that Winston Yamashita had (also) made this point (March 18, 1997) on the "free" pages, miscellaneous section, of Stanford Wong's web site.

Table 3
f^* versus a

a	f^*	a	f^*	a	f^*
0	.0200	1/3	.0120	.7	.0040
.1	.0178	.4	.0103	.8	.0024
.2	.0154	.5	.0080	.9	.0011
.3	.0128	.6	.0059	1.0	.0000

For this example, we find the new f^* for a given value of a, $0 < a < 1$, by solving $g'(f) = 0$. A value of $a = 1/3$, for instance, corresponds to a bet of $1/3$ unit on Y and 1 unit on X, a betting range of 3 to 1. The overall expectation is .01. Calculation shows $f^* = .012001$. Table 3 shows how f^* varies with a.

To understand why Cases 1 and 2 have different f^*, look first at Equation (5.1). The part of $g(f)$ corresponding to the unfavorable situations is fixed when f_0 is fixed. Only the part of $g(f)$ corresponding to the favorable situations is affected by varying f. Thus we maximize $g(f)$ by maximizing it over just the favorable situations. Whatever the result, it is then reduced by a fixed quantity, the part of g containing f_0. On the other hand, in Equation (5.2) both parts of $g(f)$ are affected when f varies, because the fraction af used for unfavorable situations bears the constant ratio a to the fraction f used in favorable situations. Now the first term, for the favorable situations, has a maximum at $f = .02$, and is approximately "flat" nearby. But the second term, for the unfavorable situations, is negative and decreasing moderately rapidly at $f = .02$. Therefore, it we reduce f somewhat, this term increases somewhat, while the first term decreases only very slightly. There is a net gain so we find $f^* < .02$. The greater a is, the more important is the effect of this term so the more we have to reduce f to get f^*, as Table 3 clearly shows. When there is a spectrum of favorable situations the solution is more complex and can be found through standard multivariable optimization techniques.

The more complex Case 2 corresponds to what the serious blackjack player is likely to need to do in practice. He will have to limit his current maximum bet to some multiple of his current minimum bet. As his bankroll increases or decreases, the corresponding bet sizes will increase or decrease proportionately.

6. Sports betting

In 1993 an outstanding young computer science Ph.D. told me about a successful sports betting system that he had developed. Upon review I was convinced. I made suggestions for minor simplifications and improvements. Then we agreed on a field test. We found a person who was extremely likely to always be regarded by the other sports bettors as a novice. I put up a test bankroll of $50,000 and we used the Kelly system to estimate our bet size.

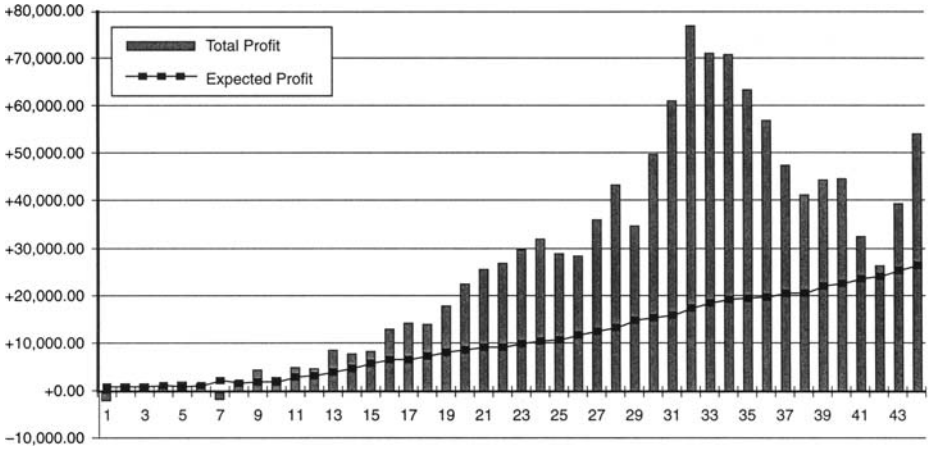


Fig. 3. Betting log Type 2 sports.

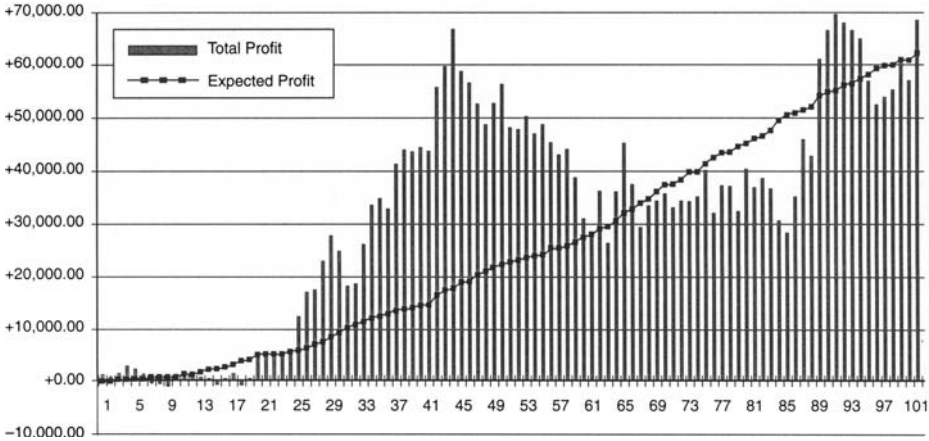


Fig. 4. Betting log Type 1 sports.

We bet on 101 days in the first four and a half months of 1994. The system works for various sports. The results appear in Figures 3 and 4. After 101 days of bets, our $50,000 bankroll had a profit of $123,000, about $68,000 from Type 1 sports and about $55,000 from Type 2 sports. The expected returns are shown as about $62,000 for Type 1 and about $27,000 for Type 2. One might assign the additional $34,000 actually won to luck. But this is likely to be at most partly true because our expectation estimates from the model were deliberately chosen to be conservative. The reason is that using too large an f^* and overbetting is much more severely penalized than using too small an f^* and underbetting.

Though $123,000 is a modest sum for some, and insignificant by Wall Street standards, the system performed as predicted and passed its test. We were never more than a few thousand behind. The farthest we had to invade our bankroll to place bets was about $10,000.

Our typical expectation was about 6% so our total bets ("action") were about $2,000,000 or about $20,000 per day. We typically placed from five to fifteen bets a day and bets ranged from a few hundred dollars to several thousand each, increasing as our bankroll grew.

Though we had a net win, the net results by casino varied by chance from a substantial loss to a large win. Particularly hard hit was the "sawdust joint" Little Caesar's. It "died" towards the end of our test and I suspect that sports book losses to us may have expedited its departure.

One feature of sports betting which is of interest to Kelly users is the prospect of betting on several games at once. This also arises in blackjack when (a) a player bets on multiple hands or (b) two or more players share a common bankroll. The standard techniques readily solve such problems. We illustrate with:

Example 6.1. Suppose we bet simultaneously on two independent favorable coins with betting fractions f_1 and f_2 and with success probabilities p_1 and p_2, respectively. Then the expected growth rate is given by

$$g(f_1, f_2) = p_1 p_2 \ln(1 + f_1 + f_2) + p_1 q_2 \ln(1 + f_1 - f_2) + q_1 p_2 \ln(1 - f_1 + f_2) + q_1 q_2 \ln(1 - f_1 - f_2).$$

To find the optimal f_1^* and f_2^* we solve the simultaneous equations $\partial g/\partial f_1 = 0$ and $\partial g/\partial f_2 = 0$. The result is

$$\begin{aligned} f_1 + f_2 &= \frac{p_1 p_2 - q_1 q_2}{p_1 p_2 + q_1 q_2} \equiv c, \\ f_1 - f_2 &= \frac{p_1 q_2 - q_1 p_2}{p_1 q_2 + q_1 p_2} \equiv d, \\ f_1^* &= (c + d)/2, \qquad f_2^* = (c - d)/2. \end{aligned} \tag{6.1}$$

These equations pass the symmetry check: interchanging 1 and 2 throughout maps the equation set into itself.

An alternate form is instructive. Let $m_i = p_i - q_i$, $i = 1, 2$ so $p_i = (1 + m_i)/2$ and $q_i = (1 - m_i)/2$. Substituting in (6.1) and simplifying leads to:

$$\begin{aligned} c &= \frac{m_1 + m_2}{1 + m_1 m_2}, \qquad d = \frac{m_1 - m_2}{1 - m_1 m_2}, \\ f_1^* &= \frac{m_1(1 - m_2^2)}{1 - m_1^2 m_2^2}, \qquad f_2^* = \frac{m_2(1 - m_1^2)}{1 - m_1^2 m_2^2} \end{aligned} \tag{6.2}$$

which shows clearly the factors by which the f_i^* are each reduced from m_i^*. Since the m_i are typically a few percent, the reduction factors are typically very close to 1.

In the special case $p_1 = p_2 = p$, $d = 0$ and $f^* = f_1^* = f_2^* = c/2 = (p - q)/(2(p^2 + q^2))$. Letting $m = p - q$ this may be written $f^* = m/(1 + m^2)$ as the optimal fraction to bet on each coin simultaneously, compared to $f^* = m$ to bet on each coin sequentially.

Our simultaneous sports bets were generally on different games and typically not numerous so they were approximately independent and the appropriate fractions were only moderately less than the corresponding single bet fractions. Question: Is this always true for independent simultaneous bets? Simultaneous bets on blackjack hands at different tables are independent but at the same table they have a pairwise correlation that has been estimated at .5 (Griffin, 1979, p. 142). This should substantially reduce the Kelly fraction per hand. The blackjack literature discusses approximations to these problems. On the other hand, correlations between the returns on securities can range from nearly -1 to nearly 1. An extreme correlation often can be exploited to great advantage through the techniques of "hedging". The risk averse investor may be able to acquire combinations of securities where the expectations add and the risks tend to cancel. The optimal betting fraction may be very large.

The next example is a simple illustration of the important effect of covariance on the optimal betting fraction.

Example 6.2. We have two favorable coins as in the previous example but now their outcomes need not be independent. For simplicity assume the special case where the two bets have the same payoff distributions, but with a joint distribution as in Table 4.

Now $c + m + b = (1 + m)/2$ so $b = (1 - m)/2 - c$ and therefore $0 \leqslant c \leqslant (1 - m)/2$.

Calculation shows $\mathrm{Var}(X_i) = 1 - m^2$, $\mathrm{Cor}(X_1, X_2) = 4c - (1 - m)^2$ and $\mathrm{Cor}(X_1, X_2) = [4c - (1 - m)^2]/(1 - m^2)$. The symmetry of the distribution shows that $g(f_1, f_2)$ will have its maximum at $f_1 = f_2 = f$ so we simply need to maximize $g(f) = (c + m)\ln(1 + 2f) + c\ln(1 - 2f)$. The result is $f^* = m/(2(2c + m))$. We see that for m fixed, as c decreases from $(1 - m)/2$ and $\mathrm{cor}(X_1, X_2) = 1$, to 0 and $\mathrm{cor}(X_1, X_2) = -(1 - m)/(1 + m)$, f^* for each bet increases from $m/2$ to $1/2$, as in Table 5.

Table 4
Joint distribution of two "identical" favorable coins with correlated outcomes

X_1 :	$X_2 : 1$	-1
1	$c + m$	b
-1	b	c

Table 5
f^* increases as $\mathrm{Cor}(X_1, X_2)$ decreases

$\mathrm{Cor}(X_1, X_2)$	c	f^*
1	$(1 - m)/2$	$m/2$
0	$(1 - m^2)/4$	$m/(1 + m^2)$
$-(1 - m)/(1 + m)$	0	$1/2$

It is important to note that for an exact solution or an arbitrarily accurate numerical approximation to the simultaneous bet problem, covariance or correlation information is not enough. We need to use the entire joint distribution to construct the g function.

We stopped sports betting after our successful test for reasons including:

(1) It required a person on site in Nevada.
(2) Large amounts of cash and winning tickets had to be transported between casinos. We believed this was very risky. To the sorrow of others, subsequent events confirmed this.
(3) It was not economically competitive with our other operations.

If it becomes possible to place bets telephonically from out of state and to transfer the corresponding funds electronically, we may be back.

7. Wall street: the biggest game

To illustrate both the Kelly criterion and the size of the securities markets, we return to the study of the effects of correlation as in Example 6.2. Consider the more symmetric and esthetically pleasing pair of bets U_1 and U_2, with joint distribution given in Table 6.

Clearly $0 \leqslant a \leqslant 1/2$ and $\mathrm{Cor}(U_1, U_2) = \mathrm{Cor}(U_1, U_2) = 4a - 1$ increases from -1 to 1 as a increases from 0 to $1/2$. Finding a general solution for (f_1^*, f_2^*) appears algebraically complicated (but specific solutions are easy to find numerically), which is why we chose Example 6.2 instead. Even with reduction to the special case $m_1 = m_2 = m$ and the use of symmetry to reduce the problem to finding $f^* = f_1^* = f_2^*$, a general solution is still much less simple. But consider the instance when $a = 0$ so $\mathrm{Cor}(U_1, U_2) = -1$. Then $g(f) = \ln(1 + 2mf)$ which increases without limit as f increases. This pair of bets is a "sure thing" and one should bet as much as possible.

This is a simplified version of the classic arbitrage of securities markets: find a pair of securities which are identical or "equivalent" and trade at disparate prices. Buy the relatively underpriced security and sell short the relatively overpriced security, achieving a correlation of -1 and "locking in" a riskless profit. An example occurred in 1983. My investment partnership bought \$ 330 million worth of "old" AT&T and sold short \$332.5 million worth of when-issued "new" AT&T plus the new "seven sisters" regional telephone companies. Much of this was done in a single trade as part of what was then the largest dollar value block trade ever done on the New York Stock Exchange (December 1, 1983).

In applying the Kelly criterion to the securities markets, we meet new analytic problems. A bet on a security typically has many outcomes rather than just a few, as in

Table 6
Joint distribution of U_1 and U_2

U_1 :	$U_2 : m_2 + 1$	$m_2 - 1$
$m_1 + 1$	a	$1/2 - a$
$m_1 - 1$	$1/2 - a$	a

most gambling situations. This leads to the use of continuous instead of discrete probability distributions. We are led to find f to maximize $g(f) = E\ln(1 + fX) = \int \ln(1 + fx)\,\mathrm{d}P(x)$ where $P(x)$ is a probability measure describing the outcomes. Frequently the problem is to find an optimum portfolio from among n securities, where n may be a "large" number. In this case x and f are n-dimension vectors and fx is their scalar product. We also have constraints. We always need $1 + fx > 0$ so $\ln(\cdot)$ is defined, and $\sum f_i = 1$ (or some $c > 0$) to normalize to a unit (or to a $c > 0$) investment. The maximization problem is generally solvable because $g(f)$ is concave. There may be other constraints as well for some or all i such as $f_i \geqslant 0$ (no short selling), or $f_i \leqslant M_i$ or $f_i \geqslant m_i$ (limits amount invested in ith security), or $\sum |f_i| \leqslant M$ (limits total leverage to meet margin regulations or capital requirements). Note that in some instances there is not enough of a good bet or investment to allow betting the full f^*, so one is forced to underbet, reducing somewhat both the overall growth rate and the risk. This is more a problem in the gaming world than in the much larger securities markets. More on these problems and techniques may be found in the literature.

7.1. Continuous approximation

There is one technique which leads rapidly to interesting results. Let X be a random variable with $P(X = m + s) = P(X = m - s) = .5$. Then $E(X) = m$, $\mathrm{Var}(X) = s^2$. With initial capital V_0, betting fraction f, and return per unit of X, the result is

$$V(f) = V_0\big(1 + (1 - f)r + fX\big) = V_0\big(1 + r + f(X - r)\big),$$

where r is the rate of return on the remaining capital, invested in, e.g., Treasury bills. Then

$$\begin{aligned} g(f) &= E\big(G(f)\big) = E\big(\ln\big(V(f)/V_0\big)\big) = E\ln\big(1 + r + f(X - r)\big) \\ &= .5\ln\big(1 + r + f(m - r + s)\big) + .5\ln\big(1 + r + f(m - r - s)\big). \end{aligned}$$

Now subdivide the time interval into n equal independent steps, keeping the same drift and the same total variance. Thus m, s^2 and r are replaced by m/n, s^2/n and r/n, respectively. We have n independent X_i, $i = 1, \ldots, n$, with

$$P\big(X_i = m/n + sn^{-1/2}\big) = P\big(X_i = m/n - sn^{-1/2}\big) = .5.$$

Then

$$V_n(f)/V_0 = \prod_{i=1}^{n}\big(1 + (1 - f)r + fX_i\big).$$

Taking $E(\log(\cdot))$ of both sides gives $g(f)$. Expanding the result in a power series leads to

$$g(f) = r + f(m - r) - s^2 f^2/2 + \mathrm{O}\big(n^{-1/2}\big) \tag{7.1}$$

where $\mathrm{O}(n^{-1/2})$ has the property $n^{1/2}\mathrm{O}(n^{-1/2})$ is bounded as $n \to \infty$. Letting $n \to \infty$ in (7.1) we have

$$g_\infty(f) \equiv r + f(m - r) - s^2 f^2/2. \tag{7.2}$$

The limit $V \equiv V_\infty(f)$ of $V_n(f)$ as $n \to \infty$ corresponds to a log normal diffusion process, which is a well-known model for securities prices. The "security" here has instantaneous drift rate m, variance rate s^2, and the riskless investment of "cash" earns at an instantaneous rate r. Then $g_\infty(f)$ in (7.2) is the (instantaneous) growth rate of capital with investment or betting fraction f. There is nothing special about our choice of the random variable X. Any bounded random variable with mean $E(X) = m$ and variance $\mathrm{Var}(X) = s^2$ will lead to the same result. Note that f no longer needs to be less than or equal to 1. The usual problems, with $\log(\cdot)$ being undefined for negative arguments, have disappeared. Also, $f < 0$ causes no problems. This simply corresponds to selling the security short. If $m < r$ this could be advantageous. Note further that the investor who follows the policy f must now adjust his investment "instantaneously". In practice this means adjusting in tiny increments whenever there is a small change in V. This idealization appears in option theory. It is well known and does not prevent the practical application of the theory (Black and Scholes, 1973). Our previous growth functions for finite sized betting steps were approximately parabolic in a neighborhood of f^* and often in a range up to $0 \leqslant f \leqslant 2f^*$, where also often $2f^* \doteq f_c$. Now with the limiting case (7.2), $g_\infty(f)$ is exactly parabolic and very easy to study.

Lognormality of $V(f)/V_0$ means $\log(V(f)/V_0)$ is $N(M, S^2)$ distributed, with mean $M = g_\infty(f)t$ and variance $S^2 = \mathrm{Var}(G_\infty(f))t$ for any time t. From this we can determine, for instance, the expected capital growth and the time t_k required for $V(f)$ to be at least k standard deviations above V_0. First, we can show by our previous methods that $\mathrm{Var}(G_\infty(f)) = s^2 f^2$, hence $\mathrm{Sdev}(G_\infty(f)) = sf$. Solving $t_k g_\infty = k t_k^{1/2} \mathrm{Sdev}(G_\infty(f))$ gives $t_k g_\infty^2$ hence the expected capital growth $t_k g_\infty$, from which we find t_k. The results are summarized in Equations (7.3).

$$
\begin{aligned}
&f^* = (m-r)/s^2, \qquad g_\infty(f) = r + f(m-r) - s^2 f^2/2,\\
&g_\infty(f^*) = (m-r)^2/2s^2 + r,\\
&\mathrm{Var}\big(G_\infty(f)\big) = s^2 f^2, \qquad \mathrm{Sdev}\big(G_\infty(f)\big) = sf,\\
&t_k g_\infty(f) = k^2 s^2 f^2 / g_\infty,\\
&t_k = k^2 s^2 f^2 / g_\infty^2.
\end{aligned}
\tag{7.3}
$$

Examination of the expressions for $t_k g_\infty(f)$ and t_k show that each one increases as f increases, for $0 \leqslant f < f_+$ where f_+ is the positive root of $s^2 f^2/2 - (m-r)f - r = 0$ and $f_+ > 2f^*$.

Comment: The capital asset pricing model (CAPM) says that the market portfolio lies on the Markowitz efficient frontier E in the (s, m) plane at a (generally) unique point $P = (s_0, m_0)$ such that the line determined by P and $(s = 0, m = r)$ is tangent to E (at P). The slope of this line is the Sharpe ratio $S = (m_0 - r_0)/s_0$ and from (7.3) $g_\infty(f^*) = S^2/2 + r$ so the maximum growth rate $g_\infty(f^*)$ depends, for fixed r, only on the Sharpe ratio. (See Quaife (1995).) Again from (7.3), $f^* = 1$ when $m = r + s^2$ in which case the Kelly investor will select the market portfolio without borrowing or lending. If $m > r + s^2$ the Kelly investor will use leverage and if $m < r + s^2$ he will

invest partly in T-bills and partly in the market portfolio. Thus the Kelly investor will dynamically reallocate as f^* changes over time because of fluctuations in the forecast m, r and s^2, as well as in the prices of the portfolio securities.

From (7.3), $g_\infty(1) = m - s^2/2$ so the portfolios in the (s, m) plane satisfying $m - s^2/2 = C$, where C is a constant, all have the same growth rate. In the continuous approximation, the Kelly investor appears to have the utility function $U(s, m) = m - s^2/2$. Thus, for any (closed, bounded) set of portfolios, the best portfolios are exactly those in the subset that maximizes the one parameter family $m - s^2/2 = C$. See Kritzman (1998), for an elementary introduction to related ideas.

Example 7.1. The long run revisited. For this example let $r = 0$. Then the basic equations (7.3) simplify to

$$\begin{aligned} &r = 0: \quad f^* = m/s^2, \quad g_\infty(f) = mf - s^2 f^2/2, \\ &g_\infty(f^*) = m^2/2s^2, \\ &\operatorname{Var}\big(G_\infty(f)\big) = s^2 f^2, \qquad \operatorname{Sdev}\big(G_\infty(f)\big) = sf. \end{aligned} \tag{7.4}$$

How long will it take for $V(f^*) \geqslant V_0$ with a specified probability? How about $V(f^*/2)$? To find the time t needed for $V(f) \geqslant V_0$ at the k standard deviations level of significance ($k = 1$, $P = 84\%$; $k = 2$, $P = 98\%$, etc.) we solve for $t \equiv t_k$:

$$tg_\infty(f) = kt^{1/2} \operatorname{Sdev}\big(G_\infty(f)\big). \tag{7.5}$$

We get more insight by normalizing all f with f^*. Setting $f = cf^*$ throughout, we find when $r = 0$

$$\begin{aligned} &r = 0: \quad f^* = m/s^2, \quad f = cm/s^2, \\ &g_\infty(cf^*) = m^2\big(c - c^2/2\big)/s^2, \\ &\operatorname{Sdev}\big(G_\infty(cf^*)\big) = cm/s, \\ &tg_\infty(cf^*) = k^2 c/(1 - c/2), \\ &t(k, cf^*) = k^2 s^2/\big(m^2(1 - c/2)^2\big). \end{aligned} \tag{7.6}$$

Equations (7.6) contain a remarkable result: $V(f) \geqslant V_0$ at the k standard deviation level of significance occurs when expected capital growth $tg_\infty = k^2 c/(1 - c/2)$ and this result is *independent of m and s*. For $f = f^*$ ($c = 1$ in (7.6)), this happens for $k = 1$ at $tg_\infty = 2$ corresponding to $V = V_0 e^2$ and at $k = 2$ for $tg_\infty = 8$ corresponding to $V = V_0 e^8$. Now $e^8 \doteq 2981$ and at a 10% annual (instantaneous) growth rate, it takes 80 years to have a probability of 98% for $V \geqslant V_0$. At a 20% annual instantaneous rate it takes 40 years. However, for $f = f^*/2$, the number for $k = 1$ and 2 are $tg_\infty = 2/3$ and $8/3$, respectively, just $1/3$ as large. So the waiting times for Prob($V \geqslant V_0$) to exceed 84% and 98% become 6.7 years and 26.7 years, respectively, and the expected growth rate is reduced to $3/4$ of that for f^*.

Comment: Fractional Kelly versus Kelly when $r = 0$

From Equations (7.6) we see that $g_\infty(cf^*)/g_\infty(f^*) = c(2-c)$, $0 \leqslant c < \infty$, showing how the growth rate relative to the maximum varies with c. The relative risk $\text{Sdev}(G_\infty(cf^*))/\text{Sdev}(G_\infty(f^*)) = c$ and the relative time to achieve the same expected total growth is $1/c(2-c)$, $0 < c < 2$. Thus the relative "spread" for the same expected total growth is $1/(2-c)$, $0 < c < 2$. Thus, even by choosing c very small, the spread around a given expected growth cannot be reduced by $1/2$. The corresponding results are not quite as simple when $r > 0$.

7.2. *The (almost) real world*

Assume that prices change "continuously" (no "jumps"), that portfolios may be revised "continuously", and that there are no transactions costs (market impact, commissions, "overhead"), or taxes (Federal, State, city, exchange, etc.). Then our previous model applies.

Example 7.2. The S&P 500 Index. Using historical data we make the rough estimates $m = .11$, $s = .15$, $r = .06$. The equations we need for $r \neq 0$ are the generalizations of (7.6) to $r \neq 0$ and $f = cf^*$, which follow from (7.3):

$$
\begin{aligned}
&cf^* = c(m-r)/s^2,\\
&g_\infty(cf^*) = \big((m-r)^2\big(c - c^2/2\big)\big)/s^2 + r,\\
&\text{Sdev}\big(G_\infty(cf^*)\big) = c(m-r)/s,\\
&tg_\infty(cf^*) = k^2c^2/\big(c - c^2/2 + rs^2/(m-r)^2\big),\\
&t(k, cf^*) = k^2c^2\big((m-r)^2/s^2\big)\big/\big(\big((m-r)^2/s^2\big)\big(c - c^2/2\big) + r\big)^2. \qquad (7.7)
\end{aligned}
$$

If we define $\widetilde{m} = m - r$, $\widetilde{G}_\infty = G_\infty - r$, $\tilde{g}_\infty = g_\infty - r$, then substitution into Equations (7.7) give Equations (7.6), showing the relation between the two sets. It also shows that examples and conclusions about $P(V_n > V_0)$ in the $r = 0$ case are equivalent to those about $P(\ln(V(t)/V_0) > rt)$ in the $r \neq 0$ case. Thus we can compare various strategies versus an investment compounding at a constant riskless rate r such as zero coupon U.S. Treasury bonds.

From Equations (7.7) and $c = 1$, we find

$$
\begin{aligned}
&f^* = 2.2\bar{2}, \qquad g_\infty(f^*) = .11\bar{5}, \qquad \text{Sdev}\big(G_\infty(f^*)\big) = .3\bar{3},\\
&tg_\infty(f^*) = .96k^2, \qquad t = 8.32k^2 \text{ years.}
\end{aligned}
$$

Thus, with $f^* = 2.2\bar{2}$, after 8.32 years the probability is 84% that $V_n > V_0$ and the expected value of $\log(V_n/V_0) = .96$ so the median value of V_n/V_0 will be about $e^{.96} = 2.61$.

With the usual unlevered $f = 1$, and $c = .45$, we find using (7.3)

$$
\begin{aligned}
&g_\infty(1) = m - s^2/2 = .09875, \qquad \text{Sdev}\big(G_\infty(1)\big) = .15,\\
&tg_\infty(1) = .23k^2, \qquad t(k, .45f^*) = 2.31k^2 \text{ years.}
\end{aligned}
$$

Writing $tg_\infty = h(c)$ in (7.7) as

$$h(c) = k^2/\big(1/c + rs^2/\big((m-r)^2c^2\big) - 1/2\big)$$

we see that the measure of riskiness, $h(c)$, increases as c increases, at least up to the point $c = 2$, corresponding to $2f^*$ (and actually beyond, up to $1 + \sqrt{1 + \frac{rs^2}{(m-r)^2}}$).

Writing $t(k, cf^*) = t(c)$ as

$$t(c) = k^2\big((m-r)^2/s^2\big)\big/\big((m-r)^2/s^2\big)(1 - c/2) + r/c^2$$

shows that $t(c)$ also increases as c increases, at least up to the point $c = 2$. Thus for smaller (more conservative) $f = cf^*$, $c \leqslant 2$, specified levels of $P(V_n > V_0)$ are reached earlier. For $c < 1$, this comes with a reduction in growth rate, which reduction is relatively small for f near f^*.

Note: During the period 1975–1997 the short term T-bill total return for the year, a proxy for r if the investor lends (i.e., $f < 1$), varied from a low of 2.90% (1993) to a high of 14.71% (1981). For details, see Ibbotson Associates, 1998 (or any later) Yearbook.

A large well connected investor might be able to borrow at broker's call plus about 1%, which might be approximated by T-bills plus 1%. This might be a reasonable estimate for the investor who borrows ($f > 1$). For others the rates are likely to be higher. For instance the prime rate from 1975–1997 varied from a low of 6% (1993) to a high of 19% (1981), according to Associates First Capital Corporation (1998).

As r fluctuates, we expect m to tend to fluctuate inversely (high interest rates tend to depress stock prices for well known reasons). Accordingly, f^* and g_∞ will also fluctuate so the long term S&P index fund investor needs a procedure for periodically re-estimating and revising f^* and his desired level of leverage or cash.

To illustrate the impact of $r_b > r$, where r_b is the investor's borrowing rate, suppose r_b in example (7.2) is $r + 2\%$ or .08, a choice based on the above cited historical values for r, which is intermediate between "good" $r_b \doteq r + 1\%$, and "poor" $r_b \doteq$ the prime rate $\doteq r + 3\%$. We replace r by r_b in Equations (7.7) and, if $f^* > 1$, $f^* = 1.3\bar{3}$, $g_\infty(f^*) = .100$, $\mathrm{Sdev}(G_\infty(f^*)) = .20$, $tg_\infty(f^*) = .4k^2$, $t = 4k^2$ years. Note how greatly f^* is reduced.

Comment: Taxes

Suppose for simplicity that all gains are subject to a constant continuous tax rate T and that all losses are subject to a constant continuous tax refund at the same rate T. Think of the taxing entities, collectively, as a partner that shares a fraction T of all gains and losses. Then Equations (7.7) become:

$$cf^* = c(m-r)/s^2(1-T),$$

$$g_\infty(cf^*) = \big((m-r)^2\big(c - c^2/2\big)\big)/s^2 + r(1-T),$$

$$\mathrm{Sdev}\big(G_\infty(cf^*)\big) = c(m-r)/s,$$

$$tg_\infty(cf^*) = k^2c^2/\big(c - c^2/2 + r(1-T)s^2/\big(m - r^2\big)\big),$$
$$t(k, cf^*) = k^2c^2\big((m-r)^2/s^2\big)\big/\Big(\big((m-r)^2/s^2\big)\big(c - c^2/2\big) + r(1-T)\Big)^2. \tag{7.7T}$$

It is interesting to see that cf^* increases by the factor $1/(1-T)$. For a high income California resident, the combined state and federal marginal tax rate is 45% so this factor is $1/.55 = 1.82$. The amplification of cf^* leads to the same growth rate as before except for a reduction by rT. The Sdev is unchanged and $t(k, cf^*)$ is increased slightly. However, as a practical matter, the much higher leverage needed with a high tax rate is typically not allowed under the margin regulation or is not advisable because the inability to continuously adjust in the real world creates dangers that increase rapidly with the degree of leverage.

7.3. The case for "fractional Kelly"

Figure 5 shows three g curves for the true m : $m_t = .5m_e$, $1.0m_e$ and $1.5m_e$, where m_e is the estimated value of m. The vertical lines and the slanting arrows illustrate the reduction in g for the three choices of: $f = .5f_e^*$, f_e^* and $1.5f_e^*$. For example with $f = .5f_e^*$ or "half Kelly", we have no loss and achieve the maximum $g = .25$, in case $m_t = .5m_e$. But if $m_t = m_e$ then $g = .75$, a loss of .25 and if $m_t = 1.5m_e$ then $g = 1.25$, a loss of 1.0, where all g are in units of $m_e^2/2s^2$. This is indicated both by LOSS_1 and LOSS_2 on the vertical line above $f/f_e^* = .5$, and by the two corresponding arrows which lead upward, and in this case to the right, from this line. A disaster occurs when $m_t = .5m_e$ but we choose $f = 1.5f_e^*$. This combines overbetting f_e^* by 50% with the overestimate of $m_e = 2m_t$. Then $g = -.75$ and we will be ruined. It is still bad to choose $f = f_e^*$ when $m_t = .5m_e$ for then $g = 0$ and we suffer increasingly wild oscillations, both up and down, around our initial capital. During a large downward oscillation experience shows that bettors will generally either quit or be eliminated by a minimum bet size requirement.

Some lessons here are:

(1) To the extent m_e is an uncertain estimate of m_t, it is wise to assume $m_t < m_e$ and to choose $f < f_e^*$ by enough to prevent $g \leqslant 0$.

Estimates of m_e in the stock market have many uncertainties and, in cases of forecast excess return, are more likely to be too high than too low. The tendency is to regress towards the mean. Securities prices follow a "non-stationary process" where m and s vary somewhat unpredictably over time. The economic situation can change for companies, industries, or the economy as a whole. Systems that worked may be partly or entirely based on data mining so m_t may be substantially less than m_e. Changes in the "rules" such as commissions, tax laws, margin regulations, insider trading laws, etc., can also affect m_t. Systems that do work attract capital, which tends to push exceptional m_t down towards average values. The drift down means $m_e > m_t$ is likely.

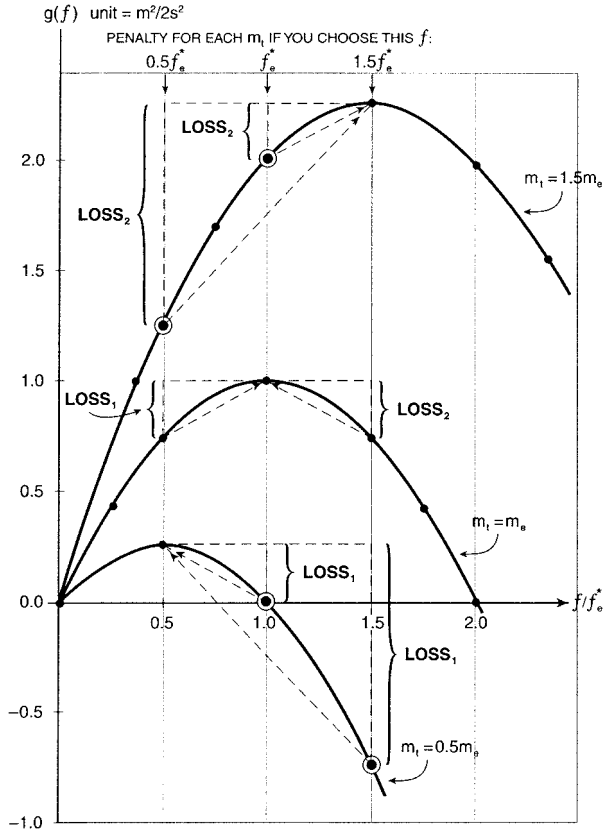


Fig. 5. Penalties for choosing $f = f_e \neq f^* = f_t$.

Sports betting has much the same caveats as the securities markets, with its own differences in detail. Rules changes, for instance, might include: adding expansion teams; the three point rule in basketball; playing overtime sessions to break a tie; changing types of bats, balls, gloves, racquets or surfaces.

Blackjack differs from the securities and sports betting markets in that the probabilities of outcomes can in principle generally be either calculated or simulated to any desired degree of accuracy. But even here m_t is likely to be at least somewhat less than m_e. Consider player fatigue and errors, calculational errors and mistakes in applying either blackjack theory or Kelly theory (e.g., calculating f^* correctly, for which some of the issues have been discussed above), effects of a fixed shuffle point, non-random shuffling, preferential shuffling, cheating, etc.

(2) Subject to (1), choosing f in the range $.5f_e^* \leqslant f < f_e^*$ offers protection against $g \leqslant 0$ with a reduction of g that is likely to be no more than 25%.

Example 7.3. The great compounder. In 1964 a young hedge fund manager acquired a substantial interest in a small New England textile company called Berkshire Hathaway. The stock traded then at 20. In 1998 it traded at 70,000, a multiple of 3500, and an annualized compound growth rate of about 27%, or an instantaneous rate of 24%. The once young hedge fund manager Warren Buffett is now acknowledged as the greatest investor of our time, and the world's second richest man. You may read about Buffett in Buffett and Clark (1997), Hagstrom (1994, 2004), Kilpatrick (1994) and Lowenstein (1995). If, as I was, you were fortunate enough to meet Buffett and identify the Berkshire opportunity, what strategy does our method suggest? Assume (the somewhat smaller drift rate) $m = .20$, $s = .15$, $r = .06$. Note: Plausible arguments for a smaller future drift rate include regression towards the mean, the increasing size of Berkshire, and risk from the aging of management. A counter-argument is that Berkshire's compounding rate has been as high in its later years as in its earlier years. However, the S&P 500 Index has performed much better in recent years so the spread between the growth rates of the Index and of Berkshire has been somewhat less. So, if we expect the Index growth rate to revert towards the historical mean, then we expect Berkshire to do so even more. From Equations (7.3) or (7.7),

$$f^* = 6.2\bar{2}, \qquad g_\infty(f^*) = .49\bar{5}, \qquad \operatorname{Sdev}\bigl(G_\infty(f^*)\bigr) = .9\bar{3},$$
$$tg_\infty(f^*) = 1.76k^2, \qquad t = 3.54k^2 \text{ years}.$$

Compare this to the unlevered portfolio, where $f = 1$ and $c = 1/6.2\bar{2} \doteq .1607$. We find:

$$f = 1, \qquad g_\infty(f) = .189, \qquad \operatorname{Sdev}\bigl(G_\infty(f)\bigr) = .15,$$
$$t_k g_\infty(f) = .119k^2, \qquad t_k = .63k^2 \text{ years}.$$

Leverage to the level $6.2\bar{2}$ would be inadvisable here in the real world because securities prices may change suddenly and discontinuously. In the crash of October, 1987, the S&P 500 index dropped 23% in a single day. If this happened at leverage of 2.0, the new leverage would suddenly be $77/27 = 2.85$ before readjustment by selling part of the portfolio. In the case of Berkshire, which is a large well-diversified portfolio, suppose we chose the conservative $f = 2.0$. Note that this is the maximum initial leverage allowed "customers" under current regulations. Then $g_\infty(2) = .295$. The values in 30 years for median V_∞/V_0 are approximately: $f = 1$, $V_\infty/V_0 = 288$; $f = 2$, $V_\infty/V_0 = 6,974$; $f = 6.2\bar{2}$, $V_\infty/V_0 = 2.86 \times 10^6$. So the differences in results of leveraging are enormous in a generation. (Note: Art Quaife reports $s = .24$ for 1980–1997. The reader is invited to explore the example with this change.)

The results of Section 3 apply directly to this continuous approximation model of a (possibly) leveraged securities portfolio. The reason is that both involve the same "dynamics", namely $\log G_n(f)$ is approximated as (scaled) Brownian motion with drift. So we can answer the same questions here for our portfolio that were answered in Section 3 for casino betting. For instance (3.2) becomes

$$\operatorname{Prob}\bigl(V(t)/V_0 \leqslant x \text{ for some } t\bigr) = x^{\wedge}\bigl(2g_\infty/\operatorname{Var}(G_\infty)\bigr) \tag{7.8}$$

where $\wedge$ means exponentiation and $0 < x < 1$. Using (7.4), for $r = 0$ and $f = f^*$, $2g_\infty / \operatorname{Var}(G_\infty) = 1$ so this simplifies to

$$\text{Prob}(\cdot) = x. \tag{7.9}$$

Compare with Example 3.3 For $0 < r < m$ and $f = f^*$ the exponent of x in (7.9) becomes $1 + 2rs^2/(m-r)^2$ and has a positive first derivative so, as r increases, $P(\cdot)$ decreases since $0 < x < 1$, tending to 0 as r tends to m, which is what we expect.

7.4. A remarkable formula

In earlier versions of this chapter the exponent in Equations (3.2), (7.8) and (7.9) were off by a factor of 2, which I had inadvertently dropped during my derivation. Subsequently Don Schlesinger posted (without details) two more general continuous approximation formulas for the $r = 0$ case on the Internet at www.bjmath.com dated June 19, 1997.

If V_0 is the initial investment and $y > 1 > x > 0$ then for f^* the probability that $V(t)$ reaches yV_0 before xV_0 is

$$\text{Prob}\big(V(t,\ f^*) \text{ reaches } yV_0 \text{ before } xV_0\big) = (1-x)/\big(1-(x/y)\big) \tag{7.10}$$

and more generally, for $f = cf^*$, $0 < c < 2$,

$$\begin{aligned}&\text{Prob}\big(V(t,\ cf^*) \text{ reaches } yV_0 \text{ before } xV_0\big)\\ &\quad = \big[1 - x\text{^}(2/c-1)\big]\big/\big[1-(x/y)\text{^}(2/c-1)\big]\end{aligned} \tag{7.11}$$

where $\wedge$ means exponentiation.

Clearly (7.10) follows from (7.11) by choosing $c = 1$. The $r = 0$ case of our Equation (7.8) follows from (7.11) and the $r = 0$ case of our Equation (7.9) follows from (7.10). We can derive a generalization of (7.11) by using the classical gambler's ruin formula (Cox and Miller, 1965, p. 31, Equation (2.0)) and passing to the limit as step size tends to zero (Cox and Miller, 1965, pp. 205–206), where we think of $\log(V(t,\ f)/V_0)$ as following a diffusion process with mean g_∞ and variance $v(G_\infty)$, initial value 0, and absorbing barriers at $\log y$ and $\log x$. The result is

$$\text{Prob}\big(V(t, cf^*) \text{ reaches } yV_0 \text{ before } xV_0\big) = [1 - x\text{^}a]/\big[1-(x/y)\text{^}a\big] \tag{7.12}$$

where $a = 2g_\infty / V(G_\infty) = 2M/V$ where M and V are the drift and variance, respectively, of the diffusion process per unit time. Alternatively, (7.12) is a simple restatement of the known solution for the Wiener process with two absorbing barriers (Cox and Miller, 1965, Example 5.5).

As Schlesinger notes, choosing $x = 1/2$ and $y = 2$ in (7.10) gives Prob($V(t,\ f^*)$ doubles before halving) $= 2/3$. Now consider a gambler or investor who focuses only on values $V_n = 2^n V_0$, $n = 0, \pm 1, \pm 2, \ldots$, multiples of his initial capital. In log space, $\log(V_n/V_0) = n \log 2$ so we have a random walk on the integer multiples of $\log 2$, where

the probability of an increase is $p = 2/3$ and of a decrease, $q = 1/3$. This gives us a convenient compact visualization of the Kelly strategy's level of risk.

If instead we choose $c = 1/2$ ("half Kelly"), Equation (7.11) gives Prob($V(t, f^*/2)$ doubles before halving) $= 8/9$ yet the growth rate $g_\infty(f^*/2) = .75 g_\infty(f^*)$ so "half Kelly" has 3/4 the growth rate but much less chance of a big loss.

A second useful visualization of comparative risk comes from Equation (7.8) which gives

$$\text{Prob}\big(V(t, cf^*)/V_0 \leqslant x \text{ for some } t\big) = x^\wedge(2/c - 1). \tag{7.13}$$

For $c = 1$ we had Prob$(\cdot) = x$ and for $c = 1/2$ we get Prob$(\cdot) = x^3$. Thus "half Kelly" has a much lessened likelihood of severe capital loss. The chance of ever losing half the starting capital is 1/2 for $f = f^*$ but only 1/8 for $f = f^*/2$. My gambling and investment experience, as well as reports from numerous blackjack players and teams, suggests that most people strongly prefer the increased safety and psychological comfort of "half Kelly" (or some nearby value), in exchange for giving up 1/4 of their growth rate.

8. A case study

In the summer of 1997 the XYZ Corporation (pseudonym) received a substantial amount of cash. This prompted a review of its portfolio, which is shown in Table 7 in the column 8/17/97. The portfolio was 54% in Biotime, ticker BTIM, a NASDAQ biotechnology company. This was due to existing and historical relationships between people in XYZ Corp. and in BTIM. XYZ's officers and directors were very knowledge-

Table 7
Statistics for logs of monthly wealth relatives, 3/31/92 through 6/30/97

		Berkshire	BioTime	SP500	T-bills
Monthly	Mean	.0264	.0186	.0146	.0035
	Standard deviation	.0582	.2237	.0268	.0008
Annual	Mean	.3167	.2227	.1753	.0426
	Standard deviation	.2016	.7748	.0929	.0028
Monthly	Covariance	.0034	−.0021	.0005	1.2E−06
			.0500	−.0001	3.2E−05
				.0007	5.7E−06
					6.7E−07
Monthly	Correlation	1.0000	−.1581	.2954	.0257
			1.0000	−.0237	.1773
				1.0000	.2610
					1.0000

able about BTIM and felt they were especially qualified to evaluate it as an investment. They wished to retain a substantial position in BTIM.

The portfolio held Berkshire Hathaway, ticker BRK, having first purchased it in 1991.

8.1. The constraints

Dr. Quaife determined the Kelly optimal portfolio for XYZ Corp. subject to certain constraints. The list of allowable securities was limited to BTIM, BRK, the Vanguard 500 (S&P 500) Index Fund, and T-bills. Being short T-bills was used as a proxy for margin debt. The XYZ broker actually charges about 2% more, which has been ignored in this analysis. The simple CAPM (capital asset pricing model) suggests that the investor only need consider the market portfolio (for which the S&P 500 is being substituted here, with well known caveats) and borrowing or lending. Both Quaife and the author were convinced that BRK was and is a superior alternative and their knowledge about and long experience with BRK supported this.

XYZ Corp. was subject to margin requirements of 50% initially and 30% maintenance, meaning for a portfolio of securities purchased that initial margin debt (money lent by the broker) was limited to 50% of the value of the securities, and that whenever the value of the account net of margin debt was less than 30% of the value of the securities owned, then securities would have to be sold until the 30% figure was restored.

In addition XYZ Corp. wished to continue with a "significant" part of its portfolio in BTIM.

8.2. The analysis and results

Using monthly data from 3/31/92 through 6/30/97, a total of 63 months, Quaife finds the means, covariances, etc. given in Table 7.

Note from Table 7 that BRK has a higher mean and a lower standard deviation than BTIM, hence we expect it to be favored by the analysis. But note also the negative correlation with BTIM, which suggests that adding some BTIM to BRK may prove advantageous.

Using the statistics from Table 7, Quaife finds the following optimal portfolios, under various assumptions about borrowing.

As expected, BRK is important and favored over BTIM but some BTIM added to the BRK is better than none.

If unrestricted borrowing were allowed it would be foolish to choose the corresponding portfolio in Table 8. The various underlying assumptions are only approximations with varying degrees of validity: Stock prices do not change continuously; portfolios can't be adjusted continuously; transactions are not costless; the borrowing rate is greater than the T-bill rate; the after tax return, if different, needs to be used; the process which generates securities returns is not stationary and our point estimates of the statistics in Table 7 are uncertain. We have also noted earlier that because "over-

Table 8
Optimal portfolio allocations with various assumptions about borrowing

Security fraction			
Security	No borrowing	50% margin	Unrestricted borrowing
Berkshire	.63	1.50	6.26
BioTime	.37	.50	1.18
S&P 500	.00	.00	12.61
T-bills	.00	−1.00	−19.04
Portfolio growth rate			
Mean	.36	.62	2.10
Standard deviation	.29	.45	2.03

betting" is much more harmful than underbetting, "fractional Kelly" is prudent to the extent the results of the Kelly calculations reflect uncertainties.

In fact, the data used comes from part of the period 1982–1997, the greatest bull market in history. We would expect returns in the future to regress towards the mean so the means in Table 7 are likely to be overestimates of the near future. The data set is necessarily short, which introduces more uncertainty, because it is limited by the amount of BTIM data. As a sensitivity test, Quaife used conservative (mean, std. dev.) values for the price relatives (not their logs) for BRK of (1.15, .20), BTIM of (1.15, 1.0) and the S&P 500 from 1926–1995 from Ibbotson (1998) of (1.125, .204) and the correlations from Table 7. The result was fractions of 1.65, .17, .18 and −1.00 respectively for BRK, BTIM, S&P 500 and T-bills. The mean growth rate was .19 and its standard deviation was .30.

8.3. The recommendation and the result

The 50% margin portfolio reallocations of Table 8 were recommended to XYZ Corp.'s board on 8/17/97 and could have been implemented at once. The board elected to do nothing. On 10/9/97 (in hindsight, a good sale at a good price) it sold some BTIM and left the proceeds in cash (not good). Finally on 2/9/98 after a discussion with both Quaife and the author, it purchased 10 BRK (thereby gaining almost $140,000 by 3/31/98, as it happened). The actual policy, led to an increase of 73.5%. What would have happened with the recommended policy with no rebalance and with one rebalance on 10/6/97? The gains would have been 117.6% and 199.4%, respectively. The gains over the suboptimal board policy were an additional $475,935 and $1,359,826, respectively.

The optimal policy displays three important features in this example: the use of leverage, the initial allocation of the portfolio, and possible rebalancing (reallocation) of the portfolio over time. Each of these was potentially important in determining the final

result. The potential impact of continuously rebalancing to maintain maximum margin is illustrated in Thorp and Kassouf (1967), Appendix A, The Avalanche Effort.

The large loss from the suboptimal policy was much more than what would have been expected because BRK and BTIM appreciated remarkably. In .62 years, BRK was up 60.4% and BTIM was up 62.9%. This tells us that—atypically—in the absence of rebalancing, the relative initial proportions of BRK and BTIM did not matter much over the actual time period. However, rebalancing to adjust the relative proportions of BRK and BTIM was important, as the actual policy's sale of some BTIM on 10/9/97 illustrated. Also, rebalancing was important for adjusting the margin level as prices, in this instance, rose rapidly.

Table 8 illustrates what we might have normally expected to gain by using 50% margin, rather than no margin. We expect the difference in the medians of the portfolio distributions to be $\$1{,}080{,}736[\exp(.62 \times .62) - \exp(.36 \times .62)] = \$236{,}316$ or 21.9% which is still large.

8.4. The theory for a portfolio of securities

Consider first the unconstrained case with a riskless security (T-bills) with portfolio fraction f_0 and n securities with portfolio fractions $f_1, \ldots, f_n$. Suppose the rate of return on the riskless security is r and, to simplify the discussion, that this is also the rate for borrowing, lending, and the rate paid on short sale proceeds. Let $C = (s_{ij})$ be the matrix such that s_{ij}, $i, j = 1, \ldots, n$, is the covariance of the ith and jth securities and $M = (m_1, m_2, \ldots, m_n)^T$ be the row vector such that m_i, $i = 1, \ldots, n$, is the drift rate of the ith security. Then the portfolio satisfies

$$\begin{aligned} &f_0 + \cdots + f_n = 1, \\ &m = f_0 r + f_1 m_1 + \cdots + f_n m_n = r + f_1(m_1 - r) + \cdots + f_n(m_n - r) \\ &\quad = r + F^T(M - R), \\ &s^2 = F^T C F \end{aligned} \tag{8.1}$$

where $F^T = (f_1, \ldots, f_n)$ and ^{T} means "transpose", and R is the column vector $(r, r, \ldots, r)^T$ of length n.

Then our previous formulas and results for one security plus a riskless security apply to $g_\infty(f_1, \ldots, f_n) = m - s^2/2$. This is a standard quadratic maximization problem. Using (8.1) and solving the simultaneous equations $\partial g_\infty / \partial f_i = 0$, $i = 1, \ldots, n$, we get

$$\begin{aligned} &F^* = C^{-1}[M - R], \\ &g_\infty(f_1^*, \ldots, f_n^*) = r + (F^*)^T C F^*/2 \end{aligned} \tag{8.2}$$

where for a unique solution we require C^{-1} to exist, i.e., $\det C \neq 0$. When all the securities are uncorrelated, C is diagonal and we have $f_i^* = (m_i - r)/s_{ii}$ or $f_i^* = (m_i - r)/s_i^2$, which agrees with Equation (7.3) when $n = 1$.

Note: BRK issued a new class of common, ticker symbol BRK.B, with the old common changing its symbol to BRK.A. One share of BRK.A can be converted to 30 shares of BRK.B at any time, but not the reverse. BRK.B has lesser voting rights and no right to assign a portion of the annual quota of charitable contributions. Both we and the market consider these differences insignificant and the A has consistently traded at about 30 times the price of the B.

If the price ratio were always exactly 30 to 1 and both these securities were included in an analysis, they would each have the same covariances with other securities, so $\det C = 0$ and C^{-1} does not exist.

If there is an initial margin constraint of q, $0 \leqslant q \leqslant 1$, then we have the additional restriction

$$|f_1| + \cdots + |f_n| \leqslant 1/q. \tag{8.3}$$

The n-dimensional subset in (8.3) is closed and bounded.

If the rate for borrowing to finance the portfolio is $r_b = r + e_b$, $e_b \geqslant 0$, and the rate paid on the short sale proceeds is $r_s = r - e_s$, $e_s \geqslant 0$, then the m in Equation (8.1) is altered. Let $x^+ = \max(x, 0)$ and $x^- = \max(0, -x)$ so $x = x^+ - x^-$ for all x. Define $f^+ = f_1^+ + \cdots + f_n^+$, the fraction of the portfolio held long. Let $f^- = f_1^- + \cdots + f_n^-$, the fraction of the portfolio held short.

Case 1. $f^+ \leqslant 1$

$$m = r + f_1(m_1 - r) + \cdots + f_n(m_n - r) - e_s f^-. \tag{8.4.1}$$

Case 2. $f^+ > 1$

$$m = r + f_1(m_1 - r) + \cdots + f_n(m_n - r) - e_b(f^+ - 1) - e_s f^-. \tag{8.4.2}$$

9. My experience with the Kelly approach

How does the Kelly-optimal approach do in practice in the securities markets? In a little-known paper (Thorp, 1971) I discussed the use of the Kelly criterion for portfolio management. Page 220 mentions that "On November 3, 1969, a private institutional investor decided to ... use the Kelly criterion to allocate its assets". This was actually a private limited partnership, specializing in convertible hedging, which I managed. A notable competitor at the time (see Institutional Investor (1998)) was future Nobel prize winner Harry Markowitz. After 20 months, our record as cited was a gain of 39.9% versus a gain for the Dow Jones Industrial Average of +4.2%. Markowitz dropped out after a couple of years, but we liked our results and persisted. What would the future bring?

Up to May 1998, twenty eight and a half years since the investment program began. The partnership and its continuations have compounded at approximately 20% annually with a standard deviation of about 6% and approximately zero correlation with the market ("market neutral"). Ten thousand dollars would, tax exempt, now be worth 18

million dollars. To help persuade you that this may not be luck, I estimate that during this period I have made about $80 billion worth of purchases and sales ("action", in casino language) for my investors. This breaks down into something like one and a quarter million individual "bets" averaging about $65,000 each, with on average hundreds of "positions" in place at any one time. Over all, it would seem to be a moderately "long run" with a high probability that the excess performance is more than chance.

10. Conclusion

Those individuals or institutions who are long term compounders should consider the possibility of using the Kelly criterion to asymptotically maximize the expected compound growth rate of their wealth. Investors with less tolerance for intermediate term risk may prefer to use a lesser function. Long term compounders ought to avoid using a greater fraction ("overbetting"). Therefore, to the extent that future probabilities are uncertain, long term compounders should further limit their investment fraction enough to prevent a significant risk of overbetting.

Acknowledgements

I thank Dr. Jerry Baesel, Professor Sid Browne, Professor Peter Griffin, Dr. Art Quaife, and Don Schlesinger for comments and corrections and to Richard Reid for posting this chapter on his website. I am also indebted to Dr. Art Quaife for allowing me to use his analysis in the case study.

This chapter has been revised and expanded since its presentation at the 10th International Conference on Gambling and Risk Taking.

Appendix A. Integrals for deriving moments of E_∞

$$I_0(a^2, b^2) = \int_0^\infty \exp\left[-\left(a^2x^2 + b^2/x^2\right)\right] \mathrm{d}x,$$
$$I_n(a^2, b^2) = \int_0^\infty x^n \exp\left[-\left(a^2x^2 + b^2/x^2\right)\right] \mathrm{d}x.$$

Given I_0 find I_2

$$\begin{aligned} I_0(a^2, b^2) &= \int_0^\infty \exp\left[-\left(a^2x^2 + b^2/x^2\right)\right] \mathrm{d}x \\ &= -\int_\infty^0 \exp\left[-\left(a^2/u^2 + b^2u^2\right)\right]\left(-\mathrm{d}u/u^2\right) \end{aligned}$$

where $x = 1/u$ and $\mathrm{d}x = -\mathrm{d}u/u^2$ so

$$I_0\big(a^2, b^2\big) = \int_0^\infty x^{-2} \exp\big[-\big(b^2x^2 + a^2/x^2\big)\big] = I_{-2}\big(b^2, a^2\big),$$

hence

$$I_{-2}\big(a^2, b^2\big) = I_0\big(b^2, a^2\big) = \frac{\sqrt{\pi}}{2 \mid b \mid}\mathrm{e}^{-2|ab|},$$

$$I_0 = \int_0^\infty \exp\big[-\big(a^2x^2 + b^2/x^2\big)\big]\,\mathrm{d}x = U \cdot V|_0^\infty - \int_0^\infty V\,\mathrm{d}U \tag{A.1}$$

where $U = \exp[\,\cdot\,]$, $\mathrm{d}V = \mathrm{d}x$, $\mathrm{d}U = (\exp[\,\cdot\,])(-2a^2x + 2b^2x^{-3})$ and $V = x$ so

$$\begin{aligned} I_0 &= \exp\big[-\big(a^2x^2 + b^2/x^2\big)\big] \cdot x|_0^\infty \\ &\quad - \int_0^\infty \big(-2a^2x^2 + 2b^2/x^2\big)\exp\big[-\big(a^2x^2 + b^2/x^2\big)\big]\,\mathrm{d}x \\ &= 2a^2 I_2\big(a^2, b^2\big) - 2b^2 I_{-2}\big(a^2, b^2\big). \end{aligned}$$

Hence:

$$I_0\big(a^2, b^2\big) = 2a^2 I_2\big(a^2, b^2\big) - 2b^2 I_{-2}\big(a^2, b^2\big)$$

and $I_{-2}(a^2, b^2) = I_0(b^2, a^2)$ so substituting and solving for I_2 gives

$$I_2\big(a^2, b^2\big) = \frac{1}{2a^2}\big\{I_0\big(a^2, b^2\big) + 2b^2 I_0\big(b^2, a^2\big)\big\}.$$

Comments.

(1) We can solve for all even n by using I_0, I_{-2} and I_2, and integration by parts.

(2) We can use the indefinite integral J_0 corresponding to I_0, and the previous methods, to solve for J_{-2}, J_2, and then for all even n. Since

$$I_0\big(a^2, b^2\big) = \frac{\sqrt{\pi}}{2|a|}\mathrm{e}^{-2|ab|}$$

then

$$I_{-2}\big(a^2, b^2\big) = \frac{\sqrt{\pi}}{2|b|}\mathrm{e}^{-2|ab|} \quad \text{and}$$

$$I_2\big(a^2, b^2\big) = \frac{1}{2a^2}\left\{\frac{\sqrt{\pi}}{2|a|} + 2b^2\frac{\sqrt{\pi}}{2|b|}\right\}\mathrm{e}^{-2|ab|} = \frac{\sqrt{\pi}}{4a^2}\mathrm{e}^{-2|ab|}\big\{1/|a| + 2|b|\big\}.$$

Appendix B. Derivation of formula (3.1)

This is based on a note from Howard Tucker. Any errors are mine.

From the paper by Paranjape and Park, if $x(t)$ is standard Brownian motion, if $a \neq 0$, $b > 0$,

$$P\big(X(t) \leqslant at + b,\ 0 \leqslant t \leqslant T \mid X(T) = s\big) = \begin{cases} 1 - \exp\left\{-\dfrac{2b}{T}(aT + B - s)\right\} & \text{if } s \leqslant aT + b, \\ 0 & \text{if } s > aT + B. \end{cases}$$

Write this as:

$$P\big(X(t) \leqslant at + b,\ 0 \leqslant t \leqslant T \mid X(T)\big) \overset{\text{a.s.}}{=} 1 - \exp\left\{-2b\big(aT + b - X(T)\big)\frac{1}{T}\right\} \quad \text{if } X(T) \leqslant aT + b.$$

Taking expectations of both sides of the above, we get

$$\begin{aligned} &P\big(X(t) \leqslant at + b,\ 0 \leqslant t \leqslant T\big) \\ &= \int_{-\infty}^{aT+b} \big(1 - \mathrm{e}^{-2b(aT+b-s)1/T}\big)\frac{1}{\sqrt{2\pi T}}\mathrm{e}^{-s^2/2T}\,\mathrm{d}s \\ &= \frac{1}{\sqrt{2\pi T}}\int_{-\infty}^{aT+b} \mathrm{e}^{-s^2/2T}\,\mathrm{d}s - \frac{\mathrm{e}^{-2ab}}{\sqrt{2\pi T}}\int_{-\infty}^{aT+b} \mathrm{e}^{-(s-2b)^2/2T}\,\mathrm{d}s. \end{aligned}$$

Hence

$$\begin{aligned} &P(X \text{ going above line } at + b \text{ during } [0, T]) = 1 - \text{previous probability} \\ &= \frac{1}{\sqrt{2\pi T}}\int_{aT+b}^{\infty} \mathrm{e}^{-s^2/2T}\,\mathrm{d}s + \mathrm{e}^{-2ab} \cdot \frac{1}{\sqrt{2\pi T}}\int_{-\infty}^{aT-b} \mathrm{e}^{-u^2/2T}\,\mathrm{d}u, \\ &\quad \text{where } u = s - 2b. \end{aligned} \tag{B.1}$$

Now, when $a = 0$, $b > 0$,

$$P\Big[\sup_{0 \leqslant t \leqslant T} X(t) \geqslant b\Big] = \sqrt{\frac{2}{\pi T}}\int_b^{\infty} \mathrm{e}^{-v^2/2T}\,\mathrm{d}v,$$

which agrees with a known formula (see, e.g., p. 261 of Tucker (1967)). In the case $a > 0$, when $T \to \infty$, since $\sqrt{T}/T \to 0$ and $\sqrt{T} = s.d.$ of $X(T)$, the first integral $\to 0$, the second integral $\to 1$, and $P(X$ ever rises above line $at + b) = \mathrm{e}^{-2ab}$. Similarly, in the case $a < 0$, P(ever rises above line $at + b) = 1$.

The theorem it comes from is due to Sten Malmquist, On certain confidence contours for distribution functions, Ann. Math. Stat. 25 (1954), pp. 523–533. This theorem is stated in S.R. Paranjape and C. Park, Distribution of the supremum of the two-parameter Yeh–Wiener process on the boundary, J. Appl. Prob. 10 (1973).

Letting $\alpha = a\sqrt{T}$, $\beta = b/\sqrt{T}$, formula (B.1) becomes

$$P(\cdot) = N(-\alpha - \beta) + \mathrm{e}^{-2\alpha\beta}N(\alpha - \beta) \text{ where } \alpha, \beta > 0 \quad \text{or}$$

$$P\big(X(t) \leqslant at + b,\ 0 \leqslant t \leqslant T\big) = 1 - P(\cdot) = N(\alpha + \beta) - \mathrm{e}^{-2\alpha\beta} N(\alpha - \beta)$$

for the probability the line is never surpassed. This follows from:

$$\frac{1}{\sqrt{2\pi T}} \int_{aT+b}^{\infty} \mathrm{e}^{-s^2/2T}\,\mathrm{d}s = \frac{1}{\sqrt{2\pi}} \int_{a\sqrt{T}+b/\sqrt{T}}^{\infty} \mathrm{e}^{-x^2/2}\,\mathrm{d}x = N(-\alpha - \beta) \quad \text{and}$$

$$\frac{1}{\sqrt{2\pi T}} \int_{-\infty}^{aT-b} \mathrm{e}^{-u^2/2T}\,\mathrm{d}u = N(\alpha - \beta)$$

where $s = aT + b$, $x = s/\sqrt{T} = a\sqrt{T} + b/\sqrt{T}$, $\alpha = a\sqrt{T}$ and $\beta = b/\sqrt{T}$.

The formula becomes:

$$\begin{aligned} &P\big(\sup\big[X(t) - (at + b)\big] \geqslant 0:\ 0 \leqslant t \leqslant T\big) \\ &\quad = N(-\alpha - \beta) + \mathrm{e}^{-2ab} N(\alpha - \beta) \\ &\quad = N(-\alpha - \beta) + \mathrm{e}^{-2\alpha\beta} N(\alpha - \beta), \quad \alpha, \beta > 0. \end{aligned}$$

Observe that

$$\begin{aligned} P(\cdot) &< N(-\alpha - \beta) + N(\alpha - \beta) = \big\{1 - N(\alpha + \beta)\big\} + N(\alpha - \beta) \\ &= \int_{-\infty}^{\alpha-\beta} \alpha(x)\,\mathrm{d}x + \int_{\alpha+\beta}^{\infty} \alpha(x)\,\mathrm{d}x < 1 \end{aligned}$$

as it should be.

Appendix C. Expected time to reach goal

Reference: Handbook of Mathematical Functions, Abramowitz and Stegun, Editors, N.B.S. Applied Math. Series 55, June 1964.

P. 304, 7.4.33 gives with $\operatorname{erf} z \equiv \frac{2}{\sqrt{\pi}} \int_0^z \mathrm{e}^{-t^2}\,\mathrm{d}t$ the integral:

$$\begin{aligned} &\int \exp\big\{-\big(a^2x^2 + b^2/x^2\big)\big\}\,\mathrm{d}x \\ &\quad = \frac{\sqrt{\pi}}{4a}\big[\mathrm{e}^{2ab}\operatorname{erf}(ax + b/x) + \mathrm{e}^{-2ab}\operatorname{erf}(ax - b/x)\big] + C, \quad a \neq 0. \end{aligned} \tag{C.1}$$

Now the left side is >0 so for real a, we require $a > 0$ otherwise the right side is <0, a contradiction.

We also note that p. 302, 7.4.3. gives

$$\int_0^{\infty} \exp\big\{-\big(at^2 + b/t^2\big)\big\}\,\mathrm{d}t = \frac{1}{2}\sqrt{\frac{\pi}{a}}\mathrm{e}^{-2\sqrt{ab}} \tag{C.2}$$

with $\Re a > 0, \Re b > 0$.

To check (C.2) v. (C.1), suppose in (C.1) $a > 0$, $b > 0$ and find $\lim_{x\to 0}$ and $\lim_{x\to\infty}$ of $\operatorname{erf}(ax + b/x)$ and $\operatorname{erf}(ax - b/x)$,

$$\lim_{x\downarrow 0+}(ax + b/x) = +\infty, \qquad \lim_{x\downarrow 0+}(ax - b/x) = -\infty,$$
$$\lim_{x\to\infty}(ax + b/x) = +\infty, \qquad \lim_{x\to\infty}(ax - b/x) = +\infty.$$

Equation (C.1) becomes

$$\frac{\sqrt{\pi}}{4a}\mathrm{e}^{-2ab}\big[\operatorname{erf}(\infty) - \operatorname{erf}(-\infty)\big] = \frac{\sqrt{\pi}}{4a}\mathrm{e}^{-2ab}2\operatorname{erf}(\infty) = \frac{\sqrt{\pi}}{2a}\mathrm{e}^{-2ab}$$

since we know $\operatorname{erf}(\infty) = 1$.

In (C.2) replace a by a^2, b by b^2 to get

$$I_0\big(a^2, b^2\big) \equiv \int_0^\infty \exp\big\{-\big(a^2t^2 + b^2/t^2\big)\big\}\,\mathrm{d}t = \frac{1}{2}\frac{\sqrt{\pi}}{|a|}\mathrm{e}^{-2|ab|}$$

which is the same.

Note: if we choose the lower limit of integration to be 0 in (C.1), then we can find C:

$$\begin{aligned} 0 &= \int_0^{0^+} \exp\big\{-\big(a^2x^2 + b^2/x^2\big)\big\}\,\mathrm{d}x = \frac{\sqrt{\pi}}{4a}\big[\mathrm{e}^{2ab}\operatorname{erf}(\infty) + \mathrm{e}^{-2ab}\operatorname{erf}(-\infty)\big] + C \\ &= \frac{\sqrt{\pi}}{4a}\big[\mathrm{e}^{2ab} - \mathrm{e}^{-2ab}\big] + C. \end{aligned}$$

Whence

$$\begin{aligned} F(x) &\equiv \int_0^x \exp\big\{-\big(a^2x^2 + b^2/x^2\big)\big\}\,\mathrm{d}x \\ &= \frac{\sqrt{\pi}}{4a}\big\{\mathrm{e}^{2ab}\big[\operatorname{erf}(ax + b/x) - 1\big] + \mathrm{e}^{-2ab}\big[\operatorname{erf}(ax - b/x) + 1\big]\big\}. \end{aligned} \tag{C.3}$$

To see how (C.3) might have been discovered, differentiate:

$$\begin{aligned} F'(x) &= \exp\big\{-\big(a^2x^2 + b^2/x^2\big)\big\} \\ &= \frac{\sqrt{\pi}}{4a}\big\{\mathrm{e}^{2ab}\big(a - b/x^2\big)\operatorname{erf}'(ax + b/x) \\ &\quad + \mathrm{e}^{-2ab}\big(a + b/x^2\big)\operatorname{erf}'(ax - b/x)\big\}. \end{aligned}$$

Now $\operatorname{erf}'(z) = \frac{2}{\sqrt{\pi}}\exp(-z^2)$ so

$$\begin{aligned} \operatorname{erf}'(ax + b/x) &= \frac{2}{\sqrt{\pi}}\exp\big[-(ax + b/x)^2\big] = \frac{2}{\sqrt{\pi}}\exp\big\{-\big(a^2x^2 + b^2/x^2 + 2ab\big)\big\} \\ &= \frac{2}{\sqrt{\pi}}\mathrm{e}^{-2ab}\exp\big\{-\big(a^2x^2 + b^2/x^2\big)\big\} \end{aligned}$$

and, setting $b \leftarrow -b$,

$$\operatorname{erf}'(ax - b/x) = \frac{2}{\sqrt{\pi}} \mathrm{e}^{2ab} \exp\{-(a^2x^2 + b^2/x^2)\}$$

whence

$$\begin{aligned} F'(x) &= \frac{\sqrt{\pi}}{4a}\left\{\frac{2}{\sqrt{\pi}}(a - b/x^2) + \frac{2}{\sqrt{\pi}}(a + b/x^2)\right\} \exp\{-(a^2x^2 + b^2/x^2)\} \\ &= \frac{1}{2a}\{2a\} \exp\{-(a^2x^2 + b^2/x^2)\} = \exp\{-(a^2x^2 + b^2/x^2)\}. \end{aligned}$$

Case of interest: $a < 0$, $b > 0$.
Expect:

$$\begin{aligned} b > 0,\ a \leqslant 0 \quad &\Rightarrow \quad F(T) \uparrow 1 \text{ as } T \to \infty, \\ b > 0,\ a > 0 \quad &\Rightarrow \quad F(T) \uparrow c < 1 \text{ as } T \to \infty. \end{aligned}$$

If $b > 0$, $a = 0$:

$$F(T) = N(-\beta) + N(-\beta) = 2N\big(-b/\sqrt{T}\big) \uparrow 2N(0) = 1 \quad \text{as } T \uparrow \infty.$$

Also, as expected $F(T) \uparrow 1$ as $b \downarrow 0$.
If $b > 0$, $a < 0$: See below.
If $b > 0$, $a > 0$:

$$\begin{aligned} F(T) &= N\big(-a\sqrt{T} - b/\sqrt{T}\big) + \mathrm{e}^{-2ab} N\big(a\sqrt{T} - b/\sqrt{T}\big) \\ &\to N(-\infty) + \mathrm{e}^{-2ab} N(\infty) \\ &= \mathrm{e}^{-2ab} < 1 \quad \text{as } T \uparrow \infty. \end{aligned}$$

This is correct.

If $b = 0$: $F(T) = N(-a\sqrt{T}) + N(a\sqrt{T}) = 1$. This is correct.

Let $F(T) = P(X(t) \geqslant at + b \text{ for some } t, 0 \leqslant t \leqslant T)$ which equals $N(-\alpha - \beta) + \mathrm{e}^{-2ab} N(\alpha - \beta)$ where $\alpha = a\sqrt{T}$ and $\beta = b/\sqrt{T}$ so $ab = \alpha\beta$; we assume $b > 0$ and $a < 0$ in which case $0 \leqslant F(T) \leqslant 1$ and $\lim_{T\to\infty} F(T) = 1$, $\lim_{T\to 0} F(T) = 0$; F is a probability distribution function:

$$\begin{aligned} &\lim_{T\to 0} F(T) = N(-\infty) + \mathrm{e}^{-2ab} N(-\infty) = 0, \\ &\lim_{T\to\infty} F(T) = N(+\infty) + \mathrm{e}^{-2ab} N(-\infty) = 1. \end{aligned}$$

The density function is

$$f(T) = F'(T) = \frac{\partial}{\partial T}(-\alpha - \beta)N'(-\alpha - \beta) + \mathrm{e}^{-2ab} \frac{\partial}{\partial T}(\alpha - \beta)N'(\alpha - \beta)$$

where

$$\frac{\partial \alpha}{\partial T} = \frac{1}{2} a T^{-1/2}, \qquad \frac{\partial \beta}{\partial T} = -\frac{1}{2} b T^{-3/2},$$

$$N'(-\alpha-\beta)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(\alpha+\beta)^2/2}=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(a^2T+b^2/T+2ab)}{2}\right\},$$

$$N'(\alpha-\beta)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(\alpha-\beta)^2/2}=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(a^2T+b^2/T-2ab)}{2}\right\},$$

$$\begin{aligned}Tf(T)&=T\left(-\frac{1}{2}aT^{-1/2}+\frac{1}{2}bT^{-3/2}\right)\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-ab}\exp\left\{-\frac{(a^2T+b^2/T)}{2}\right\}\\&\quad+T\mathrm{e}^{-2ab}\left(\frac{1}{2}aT^{-1/2}+\frac{1}{2}bT^{-3/2}\right)\frac{1}{\sqrt{2\pi}}\mathrm{e}^{ab}\exp\left\{-\frac{(a^2T+b^2/T)}{2}\right\}\\&=\frac{\mathrm{e}^{-ab}}{2\sqrt{2\pi}}\left[(-aT^{+1/2}+bT^{-1/2})\exp\left\{-\frac{(a^2T+b^2/T)}{2}\right\}\right.\\&\quad\left.+\left(aT^{+1/2}+bT^{-1/2}\right)\exp\left\{-\frac{(a^2T+b^2/T)}{2}\right\}\right]\\&=\frac{b\mathrm{e}^{-ab}}{\sqrt{2\pi}}T^{1/2}\exp\left\{\frac{-(a^2T+b^2/T)}{2}\right\}.\end{aligned}$$

The expected time to the goal is

$$E_\infty=\int_0^\infty Tf(T)\,\mathrm{d}T=\frac{b\mathrm{e}^{-ab}}{\sqrt{2\pi}}\int_0^\infty T^{-1/2}\exp\left\{-\frac{(a^2T+b^2/T)}{2}\right\}\mathrm{d}T,$$

$$\left.\begin{array}{l}T^{1/2}=x\\T=x^2\\\mathrm{d}T=2x\,\mathrm{d}x\end{array}\right\}=\frac{2b\mathrm{e}^{-ab}}{\sqrt{2\pi}}\int_0^\infty\exp\left\{-\left[\left(\frac{a}{\sqrt{2}}\right)^2x^2+\left(\frac{b}{\sqrt{2}}\right)^2x^{-2}\right]\right\}\mathrm{d}x$$

$$=\frac{2b\mathrm{e}^{-ab}}{\sqrt{2\pi}}I_0\left(\left(\frac{a}{\sqrt{2}}\right)^2,\left(\frac{b}{\sqrt{2}}\right)^2\right).$$

Now

$$I_0\left(a^2,b^2\right)=\frac{\sqrt{\pi}}{2|a|}\mathrm{e}^{-2|ab|}\quad\text{so}$$

$$I_0\left(\left(\frac{a}{\sqrt{2}}\right)^2,\left(\frac{b}{\sqrt{2}}\right)^2\right)=\frac{\sqrt{\pi}}{\sqrt{2}|a|}\mathrm{e}^{-|ab|}\quad\text{whence}$$

$$E_\infty=\frac{2b\mathrm{e}^{-ab}}{\sqrt{2\pi}}\frac{\sqrt{\pi}}{\sqrt{2}|a|}\mathrm{e}^{-|ab|}=\frac{b}{|a|},\quad a<0,\ b>0.$$

Note:

$$f(T)\equiv F'(T)=\frac{b\mathrm{e}^{-ab}}{\sqrt{2\pi}}T^{-3/2}\exp\left\{\frac{-(a^2T+b^2/T)}{2}\right\}>0$$

for all a, e.g., $a<0$, so $F(T)$ is monotone increasing. Hence, since $\lim_{T\to\infty}F(T)=1$ for $a<0$ and $\lim_{T\to\infty}F(T)<1$ for $a>0$, $0\leqslant F(T)\leqslant 1$ for all T so we have more

confidence in using the formula for $a < 0$ too.

Check:	$E_\infty(a, b) \downarrow 0$ as $\downarrow -\infty$	yes,
	$E_\infty(a, b) \uparrow$ as $b \uparrow$	yes,
	$E_\infty(a, b) \uparrow$ as $\lvert a\rvert \downarrow$	yes,
note	$\lim_{a \downarrow 0^+} E_\infty(a, b) = +\infty$	as suspected.

This leads us to believe that in a fair coin toss (fair means no drift) and a gambler with finite capital, the expected time to ruin is infinite.

This is correct. Feller gives $D = z(a - z)$ as the duration of the game, where z is initial capital, ruin is at 0, and a is the goal. Then $\lim_{a \to \infty} D(a) = +\infty$.

Note: $E_\infty = b/|a|$ means the expected time is the same as the point where $aT + b$ crosses $X(t) = 0$. See Figure 2.

$$E_\infty = b/|a|, \quad a = -m/s^2, \quad b = \ln \lambda,$$

$\lambda = C/X_0 =$ normalized goal,

$$m = p \ln(1 + f) + q \ln(1 - f) \equiv g(f),$$

$$s^2 = pq\left\{\ln\left[(1 + f)/(1 - f)\right]\right\}^2,$$

Kelly fraction $f^* = p - q, \quad g(f^*) = p \ln 2p + q \ln 2q,$

For $m > 0, \quad E_\infty = (\ln \lambda) s^2 / g(f).$

Now this is the expected time in variance units. However s^2 variance units $= 1$ trial so

$$n(\lambda, f) \equiv \frac{E_\infty}{s^2} = \frac{\ln \lambda}{g(f)} = \frac{\ln \lambda}{m}$$

is the expected number of trials.

Check:	$n(\lambda, f) \uparrow \quad$ as $\lambda \uparrow$,
	$n(\lambda, f) \to \infty \quad$ as $\lambda \to \infty$,
	$n(\lambda, f) \uparrow \quad$ as $m \downarrow 0$,
	$n(\lambda, f) \to \infty \quad$ as $m \to 0$.

Now $g(f)$ has unique maximum at $g(f^*)$ where $f^* = p - q$, the "Kelly fraction", therefore $n(\lambda, f)$ has a unique minimum for $f = f^*$. Hence f^* reaches a fixed goal in least expected time in this, the continuous case, so we must be asymptotically close to least expected time in the discrete case, which this approximates increasing by well in the sense of the CLT (Central Limit Theorem) and its special case, the normal approximation to the binomial distribution. The difference here is the trials are asymmetric. The positive and negative step sizes are unequal.

References

Associates First Capital Corporation, 1998. 1997 Annual Report. Associates First Capital Corporation, Dallas, TX.

Black, F., Scholes, M., 1973. The pricing of options and corporate liabilities. Journal of Political Economy 81, 637–659.

Breiman, L., 1961. Optimal gambling systems for favorable games. In: Fourth Berkeley Symposium on Probability and Statistics, vol. I, pp. 65–78.

Browne, S., 1996. Reaching Goals by a Deadline: Continuous-Time Active Portfolio Management. Columbia University, New York.

Browne, S., 1997. The return on investment from proportional investment strategies. Advances in Applied Probability 30 (1), 216–238.

Buffett, M., Clark, D., 1997. Buffettology. Rawson Associates, Simon and Schuster, New York.

Cox, D.R., Miller, H.D., 1965. The Theory of Stochastic Processes. Wiley, New York.

Feller, W., 1966. An Introduction to Probability Theory and Its Applications, vol. I, Revised. Wiley, New York.

Griffin, P.A., 1979. The Theory of Blackjack. Huntington Press, Las Vegas. Revised 1995.

Hagstrom, R.G. Jr., 1994. The Warren Buffett Way. Wiley, New York.

Hagstrom, R.G. Jr., 2004. The Warren Buffett Way, second ed. Wiley, New York.

Hecht, R., 1995. Private correspondence.

Ibbotson Associates, 1998. Yearbook: Stocks, Bonds, Bills and Inflation (or any later edition). Ibbotson Associates, Chicago.

Institutional Investor, 1998. Ivory Tower Investing, pp. 43–55 (see p. 44), March.

Kelly, J.L., 1956. A new interpretation of information rate. Bell System Technical Journal 35, 917–926.

Kilpatrick, A., 1994. Of Permanent Value, the Story of Warren Buffet. Distributed by Southern Publishers Group, Birmingham, AL.

Kritzman, M., 1998. Risk and utility: basics. In: Bernstein, Damodaran (Eds.), Investment Management. Wiley, New York. Chapter 2.

Lowenstein, R., 1995. The Making of an American Capitalist. Random House, New York.

Markowitz, H., 1959. Portfolio Selection. Cowles Monograph, vol. 16. Wiley, New York.

McEnally, R.W., 1986. Latané's bequest: the best of portfolio strategies. Journal of Portfolio Management 12 (2), 21–30, Winter.

Quaife, A., 1995. Using the Sharpe ratio to evaluate investments. The Trans Times 4 (1), February. Trans Time Inc., Oakland, CA.

Rotando, L.M., Thorp, E.O., 1992. The Kelly criterion and the stock market. American Mathematical Monthly 99, 922–931, December.

Thorp, E.O., 1962. Beat the Dealer. Random House, New York.

Thorp, E.O., 1966. Beat the Dealer, second ed. Vintage, New York.

Thorp, E.O., 1969. Optimal gambling systems for favorable games. Review of the International Statistical Institute 37, 273–293.

Thorp, E.O., 1971. Portfolio choice and the Kelly criterion. In: Proceedings of the 1971 Business and Economics Section of the American Statistical Association, pp. 215–224.

Thorp, E.O., 1984. The Mathematics of Gambling. Lyle Stuart, Secaucus, NJ.

Thorp, E.O., Kassouf, S.T., 1967. Beat the Market. Random House, New York.

Thorp, E.O., Walden, W., 1966. A winning bet in Nevada baccarat, part I. Journal of the American Statistical Association 61, 313–328.

Tucker, H., 1967. A Graduate Course in Probability. Academic Press, San Diego, CA.

Wong, S., 1994. Professional Blackjack. Pi Yee Press, La Jolla, CA.

Chapter 10

CAPITAL GROWTH: THEORY AND PRACTICE*

LEONARD C. MACLEAN

School of Business Administration, Dalhousie University, Halifax, NS, Canada B3H 3J5

WILLIAM T. ZIEMBA

Sauder School of Business, University of British Columbia,Vancouver, BC, Canada V6T 1Z2
and
Visiting Professor of Finance, Sloan School of Management, MIT, Cambridge, MA 02138, USA

Contents

Abstract 430
Keywords 431
1. Introduction 432
2. Capital accumulation 434
  2.1. Asset prices 435
  2.2. Decision criteria 435
  2.3. Timing of decisions 436
3. Asset prices 436
  3.1. Pricing model 438
  3.2. Estimation 440
  3.3. Comparison 442
4. Growth strategies 444
  4.1. The Kelly strategy 445
  4.2. Stochastic dominance 449
  4.3. Bi-criteria problems: Fractional Kelly strategies 451
    4.3.1. Example: Investment in stocks and bonds in continuous time 455
    4.3.2. Example: Investment in stocks and bonds in discrete time 455
  4.4. Growth-security trade-off 457
    4.4.1. Example: Investing in turn-of-the-year effect 458
    4.4.2. Example: Kentucky Derby 460
    4.4.3. Example: Lotto games 461
5. Timing of decisions 463
  5.1. Control limits 463

* This research has been supported by the Natural Sciences and Engineering Research Council of Canada.

Handbook of Asset and Liability Management, Volume 1
Edited by S.A. Zenios and W.T. Ziemba
Copyright © 2006 Elsevier B.V. All rights reserved
DOI: 10.1016/S1872-0978(06)01010-6

5.1.1. Example: Investment in stocks and bonds 464
6. Legends of capital growth 465
6.1. Princeton Newport Partners 466
6.2. Kings College Chest Fund 466
6.3. Berkshire–Hathaway 468
6.4. Hong Kong Betting Syndicate 469
References 469

Abstract

In capital accumulation under uncertainty, a decision-maker must determine how much capital to invest in riskless and risky investment opportunities over time. The investment strategy yields a stream of capital, with investment decisions made so that the dynamic distribution of wealth has desirable properties. The distribution of accumulated capital to a fixed point in time and the distribution of the first passage time to a fixed level of accumulated capital are variables controlled by the investment decisions. An investment strategy which has many attractive and some not attractive properties is the growth optimal strategy, where the expected logarithm of wealth is maximized. This strategy is also referred to as the Kelly strategy. It maximizes the rate of growth of accumulated capital asymptotically. With the Kelly strategy, the first passage time to arbitrary large wealth targets is minimized, and the probability of reaching those targets is maximized. However, the strategy is very aggressive since the Arrow–Pratt risk aversion index is essentially zero. Hence, the chances of losing a substantial portion of wealth are very high, particularly if the estimates of the returns distribution are in error. In the time domain, the chances are high that the first passage to subsistence wealth occurs before achieving the established wealth goals.

This chapter is a survey of the theoretical results and practical uses of the capital growth approach. It is a companion to the chapter in this volume on the Kelly criterion by E.O. Thorp. Alternative formulations for capital growth models in discrete and continuous time are presented. Various criteria for performance and requirements for feasibility are related in an expected utility framework. Typically, there is a trade-off between growth and security with a fraction invested in an optimal growth portfolio determined by the risk aversion criteria. Models for calculating the optimal fractional Kelly investment with alternative performance criteria are formulated. The effect of estimation and modeling error on strategies and performance is discussed.

Various applications of the capital growth approach are made to futures trading, lotto games, horseracing, and the fundamental problem of asset allocation between stocks, bonds and cash. The chapter concludes with a discussion of some of the great investors and speculators, and how they used Kelly and fractional Kelly strategies in their investment programs.

Keywords

Kelly criterion, betting, long run investing, portfolio allocation, logarithmic utility, capital growth

JEL classification: C61, D81, G1

1. Introduction

The study of growth is a significant topic in economic theory. The capital raised through securities and bonds supports the economic growth of industry. At the same time, personal wealth is accumulated through the buying and selling of financial instruments. In the capital marketplace the investor trades assets, with the rates of return generating a trajectory of wealth over time.

In capital growth under uncertainty, an investor must determine how much capital to invest in riskless and risky instruments at each point in time, with a focus on the trajectory of accumulated capital to a planning horizon. Assuming prices are not affected by individual investments but rather by aggregate investments, individual decisions are made based on the projected price processes given the history of prices to date. An investment strategy which has generated considerable interest is the growth optimal or Kelly strategy, where the expected logarithm of wealth is maximized (Kelly, 1956). Researchers such as Thorp (1969, 1971, 1975), Hausch, Ziemba and Rubinstein (1981), Grauer and Hakansson (1986, 1987), Latane (1959), Mulvey and Vladimirou (1992), Rotando and Thorp (1992) and Rubinstein (1976) have used the optimal growth strategy to compute optimal portfolio weights in multi-asset and worldwide asset allocation problems. The wealth distribution of this strategy has many attractive characteristics (see, e.g., Hakansson, 1970, 1971; Markowitz, 1976, and Hakansson and Ziemba, 1995).

The Kelly or capital growth criteria maximizes the expected logarithm as its utility function period by period (Hakansson, 1971). It has desirable properties such as being myopic—todays optimal decision does not depend upon yesterdays or tomorrows data, asymptotically maximizing long run wealth almost surely, and attaining arbitrarily large wealth goals faster than any other strategy. Also in an economy with one log bettor and all other investors with essentially different strategies, the log bettor will eventually get all the economy's wealth (Hens and Schenk-Hoppe, 2005; Breiman, 1961). The drawback of log, with its essentially zero Arrow–Pratt absolute risk aversion (see Pratt (1964)), is that in the short run it is the most risky utility function one would ever consider. Since there is essentially no risk aversion, the wagers it suggests are very large and typically undiversified. Simulations show that log investors have much more final wealth most of the time than those using other strategies, but those investors can essentially go bankrupt a small percentage of the time, even facing a large number of very favorable independent investment choices (Ziemba and Hausch, 1986). One way to modify the growth-security profile is to use either ad hoc or scientifically computed (MacLean et al., 2004) fractional Kelly strategies that blend the log optimal portfolio with cash. For instance, a fractional Kelly strategy will keep accumulated capital above a specified wealth path with high probability given log normally distributed assets. This is equivalent to using a negative power utility function whose coefficient (equivalent to a risk aversion index) is determined by the fraction and vice versa. Thus one moves the risk aversion away from zero to a higher level. This results in a smoother wealth path

but has less growth. For non-lognormal asset returns distributions, the fractional Kelly is an approximate solution to the optimal risk-return trade-off.

In this chapter the traditional capital growth model and modifications to control the risk of capital loss are developed. Parameter estimation and risk control are considered in a Bayesian dynamic model where the filtration and control processes are separate. In continuous time, the Bayesian model is a generalization to the multiasset case of the random coefficients model of Browne and Whitt (1996). Given the estimated price dynamics, an investment decision is made to control the path of future wealth. Approaches to risk control of wealth at a planning horizon based on expected utility can be put in the context of stochastic dominance (Hanoch and Levy, 1969; Levy, 1998; Artzner et al., 1999; De Giorgio, 2005). This leads naturally to bi-criteria problems of risk and return, which can be solved explicitly in continuous time (MacLean, Zhao and Ziemba, 2005) or as a stochastic program in discrete time (Orgyczak and Ruszczynski, 2002; MacLean, Ziemba and Li, 2005). If wealth is to be controlled at all points along the trajectory, then the setting of wealth goals and first passage times is applicable. In continuous time, growth strategies minimize the time to wealth goals subject to a probability constraint requiring goal attainment before falling to a subsistence level (MacLean, Ziemba and Li, 2005).

In practice an investment portfolio cannot be continuously rebalanced and a realistic approach is to reconsider the investment decision at discrete points in time (Rogers, 2001). The time points can be at fixed intervals or at random times determined by the deviation of the wealth trajectory from expectations. At each rebalance time, with additional data and a change in wealth, price parameters are re-estimated and a new investment strategy is developed. This dynamic process is illustrated in Figure 1.

In the random rebalancing times approach, the excessive deviation from expectations is determined from wealth goals (MacLean, Zhao and Ziemba, 2005). Upper and lower wealth targets serve as control limits to determine random rebalance times or decision points. The control limits are significant when estimates for future price changes are substantially in error and the wealth trajectory is not proceeding as expected. The investment risk-return strategy controls risk as defined by the estimated distribution for returns on assets. The wealth limits control the risk resulting from estimation error.

Before considering the components of the dynamic system in Figure 1, the process of capital accumulation is defined.

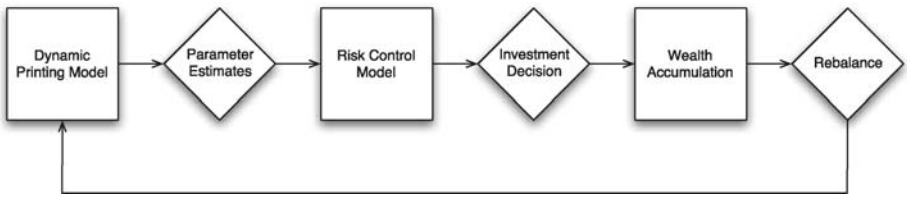


Fig. 1. Dynamic investment process.

2. Capital accumulation

Consider a competitive financial market with K assets whose prices are stochastic, dynamic processes, and a single asset whose price is non-stochastic. Let the vector of prices at time t be

$$P(t) = \big(P_0(t), P_1(t), \ldots, P_K(t)\big)', \tag{1}$$

where $P_0(t)$ is the price of the non-stochastic or risk free asset. If the prices are given at points in time t_1 and t_2, with $t_1 < t_2$, then the rate of return over that time on a unit of capital invested in asset i is

$$R_i(t_1, t_2) = \frac{P_i(t_2)}{P_i(t_1)} = 1 + r_i(t_1, t_2), \quad i = 0, \ldots, K. \tag{2}$$

In the financial market, assets are traded at points in time and the return on assets leads to the accumulation of capital for an investor. In the analysis of trading strategies, the following structure is assumed:

(a) All assets have limited liability.
(b) There are no transactions costs, taxes, or problems with indivisibility of assets.
(c) Capital can be borrowed or lent at the risk free interest rate.
(d) Short sales of all assets is allowed.

Suppose an investor has w_t units of capital at time t, and that capital is fully invested in the assets, with the proportions invested in each asset given by $x_i(t)$, $i = 0, \ldots, K$. Then an investment or trading strategy at time t is the vector process

$$X(t) = \big(x_0(t), x_1(t), \ldots, x_K(t)\big)'. \tag{3}$$

Given the investments $w_{t_1} X(t_1)$ at time t_1, the accumulated capital at time t_2 is

$$W(t_2) = w_{t_1} R'(t_1, t_2) X(t_1) = w_{t_1} \sum_{i=0}^{K} R_i(t_1, t_2) x_i(t_1). \tag{4}$$

The trajectory of returns between time t_1 and time t_2 depends on the asset, and is typically non-linear. So changing the investment strategy at points in time between t_1 and t_2 will possibly improve capital accumulation. If trades could be timed to correspond to highs and lows in prices, then greater capital would be accumulated. The history of asset prices provides information that is useful in predicting future prices, so monitoring prices and revising the investment portfolio would seem appropriate.

To consider the effect of changes in strategy, partition the time interval into n segments, with $d = \frac{t_2 - t_1}{n}$, so that the accumulated capital is monitored, and the investment strategy is possibly revised at times $t_1, t_1 + d, \ldots, t_1 + nd = t_2$. Then wealth at time t_2 is

$$W_n(t_2) = w_{t_1} \prod_{i=0}^{n-1} R'\big(t_1 + id, t_1 + (i+1)d\big) X(t_1 + id). \tag{5}$$

This wealth is

$$W_n(t_2) = w_{t_1}\left(\exp\left[\frac{1}{n}\sum_{i=0}^{n-1}\ln\bigl(R'\bigl(t_1+id, t_1+(i+1)d\bigr)X(t_1+id)\bigr)\right]\right)^n. \tag{6}$$

The exponential form highlights the *growth rate* with the strategy $X = (X(t_1), \ldots, X(t_1 + (n-1)d))$,

$$G_n(X) = \frac{1}{n}\sum_{i=0}^{n-1}\ln\bigl(R'\bigl(t_1+id, t_1+(i+1)d\bigr)X(t_1+id)\bigr). \tag{7}$$

As the partitioning of the interval gets finer, so that $d \to 0$, then monitoring and trading are continuous. Furthermore, $n \to \infty$, and when the process $V_i = \ln(R'(t_1 + id, t_1+(i+1)d)X(t_1+id))$, $i = 0, \ldots, n-1$, is ergodic in the mean, $G_n(X) \to 1+g$. If the random variables V_i, $i = 0, \ldots,$ are independent and identically distributed, then $W_n(t_2) = w_{t_1}\exp(\sum_{i=0}^{n-1} V_i)$ converges to a lognormal random variable.

The capital accumulation process is straight forward, but there are details in the setup which have significant implications for the trajectory of wealth.

2.1. Asset prices

The return on investment results from changes in asset prices over time. The investment decisions are made based on expectations about unknown future prices. The standard approach to forecasting is to propose a model for price dynamics. In practice the model is a mechanism for transforming the record of past prices into projections for future prices. There are many models that have been proposed. Probabilistic models which focus on the dynamics of the price distributions have been successful, particularly if they contain a combination of Gaussian and Poisson processes. The models approximate the transition between rates of the return with a linear combination of random processes. This format is very suited to the analysis of capital accumulation, but it is important to understand that expectations generated by the approximating model can be seriously in error. Some model flexibility is introduced by defining the parameters in the combination to be random variables and working with a Bayesian dynamic linear model. Bayes Theorem provides a natural mechanism for model revision, which would be called for when there is evidence of a significant deviation between actual and expected prices.

2.2. Decision criteria

The trajectory of accumulated capital over time depends on the returns on a unit of capital invested in assets, and the decisions on the amount of capital to invest in assets at each point in time. If the distributions for future returns are known, decisions can be based on the desired distributions of capital accumulation over time or at a planning horizon. Concepts of stochastic dominance can be used to order distributions. However, the future returns distributions are estimated with error and that risk needs to be re-

flected in investment decisions. With the volatility of accumulated capital exceeding the conditional volatility for given returns distributions, higher levels of risk aversion are appropriate. Decision models based on measures of risk and return can be used to shape the distribution of accumulated capital.

2.3. Timing of decisions

As the capital accumulation process evolves over time, information on asset returns is available. The information is important for evaluating the investment decision as well as estimating future returns distributions. There is an issue over when to react to information. It is standard practice to reconsider decisions at regular intervals in time. An alternative approach, based on concepts from statistical process control, is to set upper and lower control limits on accumulated capital, and to reconsider decisions at random times when the trajectory of capital is out of control.

These aspects of capital accumulation are developed in subsequent sections.

3. Asset prices

The returns on assets are generated by price changes. Future trading prices and therefore returns are uncertain, and characterizing the distributions for returns is the foundation for investment planning and capital growth. Over the interval (t_1, t_2), the return on asset i, $R_i(t_1, t_2)$, is the product of incremental returns:

$$R_i(t_1, t_2) = \frac{P_i(t_2)}{P_i(t_1)} = \prod_{i=1}^{n} \left(\frac{P_i(t_1 + d)}{P_i(t_1)} \right) \cdots \left(\frac{P_i(t_2)}{P_i(t_1 + (n-1)d)} \right),$$

and with the geometric form of returns, it is natural to consider the logarithm of returns when characterizing the distribution. The distribution of log returns can be defined by its moments and approximated using a small number of moments. It is known that the significance of moments diminishes with order. Chopra and Ziemba (1993) show that equal size errors in estimators for the means, variances, and covariances affect portfolio performance in the order of 20 : 2 : 1, respectively, for a Markowitz (1952, 1987) static mean variance model. (See also, the earlier studies of Kallberg and Ziemba (1981, 1984).) The performance is also a function of the investors preference for risk as reflected in the investment strategy. Low risk aversion magnifies the effect of average estimation error, and high risk aversion diminishes the effect, so for log or near log investors the 20 times error can be 80–100 times. The actual form of the assumed concave utility function is not important, its the risk aversion that matters; see Kallberg and Ziemba (1983). The performance measure in the Chopra–Ziemba study was the certainty equivalent of wealth, defined as

$$CE = u^{-1}\big(E(W)\big), \tag{8}$$

where W is accumulated capital and u is the power utility function.

Table 1
Average ratio for errors in means, variances and covariances

Risk tolerance	Errors in mean vs variance	Errors in mean vs variances
25	5.38	3.22
50	22.50	10.98
75	56.84	21.42

Source: Chopra and Ziemba (1993).

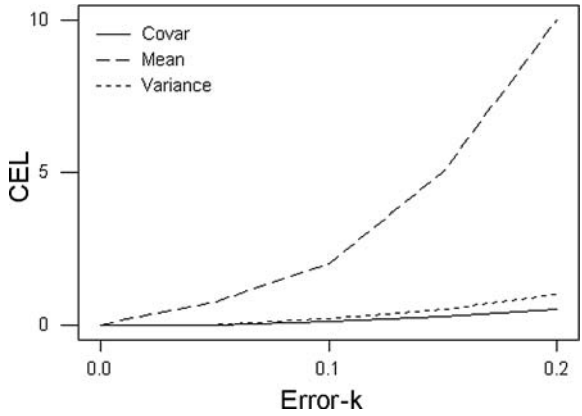


Fig. 2. Mean percent *CEL* for various errors. Source: Chopra and Ziemba (1993).

The comparison of the *CE* with the correct moment and moment contaminated by estimation error is the certainly equivalent loss

$$CEL = \left(\frac{CE_{\text{true}} - CE_{\text{cont}}}{CE_{\text{cont}}}\right)100. \tag{9}$$

The contamination was determined by the random factor $(1 + kZ)$, where $Z \propto N(0, 1)$, $k = 0.05, 0.10.0.15, 0.20$. So $Moment_{\text{cont}} = (1+kZ)Moment$. A Monte Carlo study of the effects of error (contamination) was carried out. The optimal mean-variance one period wealth was calculated for the true and contaminated moments.

Table 1 compares the ratio of the CEL for errors in the mean versus errors in variances/covariances, with various risk tolerances. Risk tolerance is the reciprocal of the Arrow–Pratt risk aversion index multiplied by $100/(1/2) = 200$.

Figure 2 shows the relative effects of errors in moment estimation as the size of the error increases.

This static analysis of the effect of errors in the moments of the returns distributions is informative, but it is important to consider the dynamics of investing. At points in time information on past returns is used to forecast future returns. Accurate estimates

for means are crucial. If investment decisions are based on expected returns, conditional on the history of returns, then capital growth is most likely. Methods of estimation for the mean rate of return in a dynamic model are considered next.

3.1. Pricing model

It is clear that getting the mean right in forecasting the returns distributions is most important. For the incremental rate of return on asset i, $\ln(\frac{P_i(t_1+jd)}{P_i(t_1+(j-1)d)})$, consider $Y_{ij}(t_1) = \ln P_i(t_1 + jd)$, and $Y_{ij}(t_1) - Y_{i(j-1)}(t_1) = \Delta Y_{ij}(t_1)$, for $i = 1, \ldots, K$. So $\ln(R_i(t_1, t_2)) = \sum_{j=1}^{n} \Delta Y_{ij}(t_1)$. In this setup, the log-returns over a time interval are approximately normal. If the mean of $\Delta Y_{ij}(t_1)$ is λ_{ij} and $\mu_i = \frac{1}{n}\sum_{j=1}^{n} \lambda_{ij}$, then the increment means can be considered as sampled from a family of means for which μ_i is the overall mean. To formalize this concept, consider the infinitesimal increments process defined by the stochastic linear equations

$$\mathrm{d}Y_0(t) = r\,\mathrm{d}t, \tag{10}$$

$$\mathrm{d}Y_i(t) = \lambda_i\,\mathrm{d}t + \delta_i\,\mathrm{d}V_i, \quad i = 1, \ldots, K, \tag{11}$$

where $\mathrm{d}V_i, i = 1, \ldots, K$, are independent Brownian motions and the rate parameters λ_i are random variables, with

$$\lambda_i = \mu_i + \beta_i Z_i, \quad i = 1, \ldots, K. \tag{12}$$

To allow for the relationship between the rates of return on assets, the Z_i, $i = 1, \ldots, K$ are correlated Gaussian variables with ρ_{ij} the instantaneous correlation between variables Z_i and Z_j. So $E(\lambda_i) = \mu_i$, and $\mathrm{cov}(\lambda_i, \lambda_j) = \sigma_{ij} = \beta_i\beta_j\rho_{ij}$.

The hierarchal linear model in (10)–(12) is a generalization of a single stock model in Browne and Whitt (1996), where the rate of return is a random variable. This model is also used in Rogers (2001) to study parameter estimation error, and in MacLean and Weldon (1996) to study random effects.

The correlated Gaussian variables can be represented in terms of independent Gaussian variables. If U_j, $j = 1, \ldots, m, m \leqslant K$, are i.i.d. standard Gaussian variables, then $Z_i = \sum_{j=1}^{m} \alpha_{ij}U_j$ and $\rho_{ij} = \sum_{k=1}^{m} \alpha_{ik}\alpha_{kj}$. In terms of the independent Gaussian variables, the model for increments is

$$\mathrm{d}Y_i(t) = \left[\mu_i + \sum_{i=1}^{m} \beta_i\alpha_{ij}U_j\right]\mathrm{d}t + \delta_i\,\mathrm{d}V_i, \quad i = 1, \ldots, K. \tag{13}$$

Typically the number of independent Gaussian variables (factors) needed to define the correlation between asset returns would be small and to have the parameters identifiable from the covariance, it is required that $m \leqslant (K-1)/2$.

Although the equation in (13) has a familiar linear form, it is important to distinguish between terms defining the rates and the volatility terms. The volatilities δ_i, $i = 1, \ldots, K$, represent the specific variance of each asset return. The asset returns

are correlated, but the correlation is generated by the factors in the expected rates of return.

There is evidence that in addition to factors in the mean rates of return, stochastic volatility factors are important for capturing certain aspects of returns distributions such as heavy tails (Chernov et al., 2002). An alternative approach to extreme returns involves adding independent shock terms to capture dramatic price changes. The dynamic equations become

$$\mathrm{d}Y_i = \lambda_i\,\mathrm{d}t + \delta_i\,\mathrm{d}V_i + \vartheta_i \mathrm{d}N_i(\pi_i),$$

where $\mathrm{d}N_i(\pi_i)$ is a Poisson process with intensity π_i, and shock size ϑ_i, $i = 1, \ldots, K$. Between shocks, the financial market is in the same regime, and is described by the equations in (10)–(12). Within a regime, the random rate model is sufficient to explain the mean, variance and covariance for returns. An approach to the model with shocks is to define a conditional model given the shocks. The conditional dynamics are in (13).

The rates of return within a regime may be dynamic, with defining equations $\mathrm{d}\lambda_i(t) = \mu_i\,\mathrm{d}t + \beta_i\,\mathrm{d}q_i$, where $\mathrm{d}q_i$, $i = 1, \ldots, K$, are correlated Brownian motions. So $\lambda_i(t) = \lambda_i(0) + \mu_i t + \beta_i\sqrt{t}Z_i$, $i = 1, \ldots, K$. The estimation methods discussed later are easily adapted to the estimation of parameters with dynamic stochastic rates of return.

The distribution of asset prices, and in particular the mean vector and covariance matrix are the objects approximated by the linear model. Returning to the pricing equations, let

$$\begin{aligned}
&Y(t) = \big(Y_1(t), \ldots, Y_K(t)\big)',\\
&\lambda = (\lambda_1, \ldots, \lambda_K)',\\
&\Delta = \operatorname{diag}\big(\delta_1^2, \ldots, \delta_K^2\big),\\
&\mu = (\mu_1, \ldots, \mu_K)',\\
&\Gamma = (\gamma_{ij}) = (\beta_i\beta_j\rho_{ij}).
\end{aligned}$$

Without loss of generality assume $Y(0) = 0$. Given (λ, Δ), the conditional distribution of log-prices at time t is

$$\big(Y(t)|\lambda, \Delta\big) \propto N(\lambda t, t\Delta). \tag{14}$$

From (12) the rate has a prior distribution

$$\lambda \propto N(\mu, \Gamma). \tag{15}$$

The marginal distribution of log-prices is

$$Y(t) \propto N(\mu t, \Sigma_t), \tag{16}$$

with

$$\Sigma_t = t^2\Gamma + t\Delta = \Gamma_t + \Delta_t. \tag{17}$$

The hierarchal Bayes model for returns characterizes asset prices in terms of means and variances/covariances. It is an important property of the model that the covariance for log-prices is partitioned into a component determined by the random drift and a component determined by the diffusion.

The model defines an evolution of asset prices, but the practical context for a model is defining future price distributions given information on past prices. At any point in time assume that information is available on the history of prices. Consider the data available at time t, $\{Y(s),\ 0 \leqslant s \leqslant t\}$, and the corresponding filtration $\Im_t^Y = \sigma\{Y(s),\ 0 \leqslant s \leqslant t\}$. The observed mean rate of return to time t is

$$\overline{Y}_t = \frac{1}{t} Y(t). \tag{18}$$

With the prior distribution for λ in (12) and the conditional distribution for $Y(t)$ in (11), the posterior distribution for the mean rate of return λ, given $\Im_t^Y$ is

$$\left(\lambda | \Im_t^Y\right) \propto N\left(\hat{\lambda}_t, \Gamma_t^0\right), \tag{19}$$

where $\hat{\lambda}_t = \mu + (I - \Delta_t \Sigma_t^{-1})(\overline{Y}_t - \mu)$ and $\Gamma_t^0 = \frac{1}{t^2}(I - \Delta_t \Sigma_t^{-1})\Delta_t$.

So the hierarchal model provides a natural filter for information on prices. The conditional distribution for incremental rates of return is modified as a trajectory of asset prices unfolds. It follows that the Bayes estimate for the mean rate of return at time t is the conditional expectation

$$\hat{\lambda}_t = E\left(\lambda | \Im_t^Y\right) = \mu + \left(I - \Lambda_t \Sigma_t^{-1}\right)\left(\overline{Y}_t - \mu\right). \tag{20}$$

The conditional expected return is the key to wealth accumulation. An investor allocating wealth across assets according to their conditional expected return eventually accumulates total market wealth (Amir et al., 2005).

The Bayes estimate in (20) and the posterior distribution depend upon unknown parameters $(\lambda, \Gamma, \Delta)$. If the parameters can be estimated from the data on past returns, then replacing $(\lambda, \Gamma, \Delta)$ with estimates $(\hat{\lambda}, \widehat{\Gamma}, \hat{\Delta})$ will provide an empirical Bayes estimate for the conditional mean rate of return.

3.2. Estimation

Assume that securities have been observed at regular intervals of width $\frac{t}{n}$ in the time period $(0, t)$. The log prices at times $\frac{(s+1)t}{n}$, given the log prices at times $\frac{st}{n}$, $s = 0, \ldots, n$, are

$$Y(s) = y(s-1) + \frac{t}{n}\lambda + \sqrt{\frac{t}{n}}\Delta^{1/2} Z. \tag{21}$$

The first order increments process generates sample rates

$$e(s) = \left(Y(s) - y(s-1)\right) \div \frac{t}{n} = \lambda + \sqrt{\frac{n}{t}}\Delta^{1/2} Z, \tag{22}$$

which are stationary with covariance

$$\Sigma_{nt} = \Gamma + \frac{n}{t}\Delta = \Gamma_{nt} + \Delta_{nt}, \tag{23}$$

and mean $E(e) = \lambda$.

From the realized trajectory of prices, the observations on log-prices at times $\frac{st}{n}$, $s = 0, \ldots, n$, are $\{Y_{is},\ i = 1, \ldots, K;\ s = 1, \ldots, n\}$. The corresponding sample rates are $\{e_{is},\ i = 1, \ldots, K;\ s = 1, \ldots, n\}$. With $e_s' = (e_{1s}, \ldots, e_{Ks})$, it follows that

$$\overline{Y}_t = \frac{1}{n}\sum_{s=1}^{n} e_s. \tag{24}$$

So $\overline{Y}_t$ is the maximum likelihood estimate of λ, given the sample rates e_s, $s = 1, \ldots, n$. Let the covariance matrix computed from the observed rates be S_{nt}, the usual estimate of Σ_{nt}. The theoretical covariance is partitioned as $\Sigma_{nt} = \Gamma_{nt} + \Delta_{nt}$, and the objective is to reproduce that partition with the sample covariance matrix. If the eigenvalues of Γ_{nt} are $\gamma_1, \ldots, \gamma_K$, and $\Delta_{nt} = \operatorname{diag}(\delta_1^2, \ldots, \delta_K^2)$, then the eigenvalues of Σ_{nt} are $\gamma_1 + \delta_1^2, \ldots, \gamma_K + \delta_K^2$. When the rank of Γ_{nt} is $m < K$, then $\gamma_{m+1} = \cdots = \gamma_K = 0$. Consider the spectral decomposition of S_{nt}, with the ordered eigenvalues $g_1, \ldots, g_K$ and the corresponding eigenvectors $l_1, \ldots, l_K$. To generate the desired sample covariance structure, choose a *truncation* value $m < K$, and define the matrices

$$D = \operatorname{diag}(g_1, \ldots, g_m), \tag{25}$$

$$L = (l_1, \ldots, l_m), \tag{26}$$

$$G_{nt} = L_{nt} D L_{nt}', \tag{27}$$

$$D_{nt} = \operatorname{diag}(S_{nt} - G_{nt}) = \operatorname{diag}(d_1, \ldots, d_K), \tag{28}$$

$$S_{nt}^* = G_{nt} + D_{nt}. \tag{29}$$

In the theoretical covariance, it is possible that the eigenvalues γ_j, $j = 1, \ldots, K$, are all positive. However, it is expected that the covariance between securities prices is generated by a small number of underlying portfolio's (factors), and therefore the number of positive eigenvalues (m) is small relative to K. In any case m is indeterminate and an arbitrary choice introduces error. Since many of the eigenvalues and eigenvectors in the above construction are discarded, the method is referred to as *truncation*.

The matrices G_{nt}, D_{nt}, S_{nt}^* are estimates of $\Gamma_{nt}, \Delta_{nt}, \Sigma_{nt}$, respectively. Therefore, $\widehat{\Gamma} = G_{nt}$ and $\hat{\Delta} = \frac{t}{n} D_{nt}$ are estimates of model parameters Γ and Δ. The estimate of Σ_t is

$$\widehat{\Sigma}_t = t^2 \widehat{\Gamma} + t\hat{\Delta} = \widehat{\Gamma}_t + \hat{\Delta}_t. \tag{30}$$

For the parameter μ, the prior mean, assumptions about the financial market can guide estimation. If it is assumed that the there is a long term equilibrium value for returns on equities, it is reasonable to say λ_i, $i = 1, \ldots, K$, have a common mean. So $\mu' =$

$(\mu, \ldots, \mu)$ and the prior mean is estimated by $\hat{\mu}_t 1$, where 1 is a vector of ones and

$$\hat{\mu}_t = \frac{1}{nK} \sum_i \sum_s e_{is}. \tag{31}$$

The *truncation estimator* for the conditional mean rate of return at time t is

$$\hat{\lambda}_{Tr} = \hat{\mu}_t 1 + \big(I - \hat{\Delta}_t \widehat{\Sigma}_t^{-1}\big)\big(\overline{Y}_t - \hat{\mu}_t 1\big). \tag{32}$$

The truncation estimator is an empirical Bayes estimator since it is in the form of the Bayes estimator, with estimates for the prior parameters. The assumption of a common prior mean could be relaxed to a common mean within asset classes, or some other grouping of securities.

An alternative empirical Bayes estimator has been developed by Jorion (1986). The prior mean is estimated by a weighted grand mean

$$\tilde{\mu}_t = 1' S_{nt}^{-1} \overline{Y}_t / \big(1' S_{nt}^{-1} 1\big). \tag{33}$$

The Bayes–Stein estimate of λ is

$$\hat{\lambda}_{\mathrm{BS}} = \tilde{\mu}_t 1 + \left(\frac{n}{\varphi + K}\right)\big(\overline{Y}_t - \tilde{\mu}_t 1\big), \tag{34}$$

with

$$\varphi = \frac{K + 2}{(\overline{Y}_t - \tilde{\mu} 1)' S_{nt}^{-1} (\overline{Y}_t - \tilde{\mu} 1)}. \tag{35}$$

Although they have similar forms, the concept behind the Bayes–Stein estimator is quite different from the truncation estimator truncation estimator. The truncation estimator adjusts the maximum likelihood estimate $\overline{Y}_t$ based on the correlation between securities prices, or equivalently the scores on the latent market factors. The prior distribution is multivariate normal and the conditional covariance (specific variance) is diagonal. The Bayes–Stein estimator shrinks all the $\overline{Y}_{it}$ toward the grand mean, based on variance reduction. In this case, the prior is univariate normal, and the conditional covariance is not diagonal.

3.3. Comparison

A comparison of the various estimators for the rate of return is provided in MacLean, Foster and Ziemba (2002) (see also MacLean, Foster and Ziemba (2005)). Two data sets for the years 1990–2002 of asset returns were used: (i) end of month prices for 24 leading stocks from the Toronto Stock Exchange (TSE); and (ii) end of month prices for 24 leading stocks from the New York Stock Exchange (NYSE). The correlation structure for prices is different for the exchanges, so a performance comparison of estimators will show the significance of structure. The percent of variance accounted for by the top 5 eigenvalues for each correlation matrix is shown in Table 2.

Table 2
Leading eigenvalues (% of variance)

	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	Total
NYSE	21.9	12.4	7.9	6.6	5.5	54.3
TSE	28.3	18.7	14.8	8.9	6.7	77.4

Source: MacLean, Foster and Ziemba (2005).

From the data on monthly closing prices for the set of 24 stocks on the Toronto Stock Exchange and the separate set of stocks on the New York Stock Exchange, the price increments (natural log of gross monthly rates of return) were computed. The mean vector and covariance matrix of the increments for each exchange were calculated, and these values were used as parameters in the dynamic model. Then trajectories of prices for 50 months were simulated and the expected rates of return were estimated by the various methods: average, truncation estimator, and Bayes–Stein estimator. For the truncation estimator the number of factors was preset at $m^* = 5$. Since the TSE data has a more compact market structure, it is expected that the truncation estimator with $m^* = 5$ will perform better in that case. This experiment was repeated 1000 times.

The standard criterion for comparing estimators $\hat{\lambda}$ of a true mean rate of return λ is based on the mean squared error matrix:

$$E\big(\hat{\lambda} - \lambda\big)\big(\hat{\lambda} - \lambda\big)'. \tag{36}$$

The risk of an estimator is defined as the trace of the MSE matrix:

$$R\big(\lambda, \hat{\lambda}\big) = \operatorname{tr} E\big(\hat{\lambda} - \lambda\big)\big(\hat{\lambda} - \lambda\big)'. \tag{37}$$

In the simulation experiment, the value $[\frac{1}{K} R(\lambda, \hat{\lambda})]^{1/2}$—root mean squared error, as well as the error for each estimator was computed. The results are shown in Table 3.

The truncation estimator has smaller mean squared error for most stocks on the NYSE and for all stocks on the TSE. As predicted, the performance is better for the stocks on the TSE.

The parameter estimation presented above has focused on the mean return, since that value has such significance in investment decisions and capital accumulation. Studies of the long term effects of improved estimates of the mean justify that emphasis (Grauer and Hakansson, 1986; Amir et al., 2005). The hierarchal model has volatility terms, and the empirical Bayes (truncation) approach also provides estimates of asset specific volatilities. The hierarchal model is in the class of stochastic volatility models. There are other approaches to forecasting returns. Most notable are CAPM and GARCH models. The model most consistent with the capital growth formulation is the hierarchal linear model, which characterizes the posterior mean rates of return given the data on past returns.

Table 3
%root mean squared error

NYSE stock	Estimator			TSE stock	Estimator		
	Tr	*Avg*	*BS*		*Tr*	*Avg*	*BS*
$S_{\text{NYSE},1}$	1.9725	2.8432	2.1462	$S_{\text{TSE},1}$	1.7668	4.5063	3.1027
$S_{\text{NYSE},2}$	4.1581	4.8305	4.2945	$S_{\text{TSE},2}$	2.7083	4.4457	3.2007
$S_{\text{NYSE},3}$	1.8415	2.9378	2.2138	$S_{\text{TSE},3}$	1.9782	4.3989	3.0300
$S_{\text{NYSE},4}$	2.2514	3.0818	2.3202	$S_{\text{TSE},4}$	1.7466	4.3747	3.0131
$S_{\text{NYSE},5}$	2.5917	2.8565	2.2977	$S_{\text{TSE},5}$	1.6950	4.5301	3.1046
$S_{\text{NYSE},6}$	2.8090	2.9188	2.2270	$S_{\text{TSE},6}$	1.8265	4.6527	3.2517
$S_{\text{NYSE},7}$	4.9898	5.3249	5.8516	$S_{\text{TSE},7}$	1.6682	4.2899	2.9744
$S_{\text{NYSE},8}$	2.1711	2.9338	2.2597	$S_{\text{TSE},8}$	1.8485	4.3125	2.9897
$S_{\text{NYSE},9}$	2.3748	3.1987	2.5019	$S_{\text{TSE},9}$	1.8717	4.8132	3.3046
$S_{\text{NYSE},10}$	2.0109	3.0979	2.4649	$S_{\text{TSE},10}$	2.0050	4.4784	3.1089
$S_{\text{NYSE},11}$	2.5031	3.0759	2.4323	$S_{\text{TSE},11}$	2.3007	4.5624	3.2394
$S_{\text{NYSE},12}$	1.9821	2.9752	2.2563	$S_{\text{TSE},12}$	2.2495	5.0386	3.4564
$S_{\text{NYSE},13}$	2.0238	2.9266	2.2242	$S_{\text{TSE},13}$	2.6598	4.3629	3.1579
$S_{\text{NYSE},14}$	1.9424	3.1798	2.5272	$S_{\text{TSE},14}$	1.9208	4.5340	3.1757
$S_{\text{NYSE},15}$	1.9684	3.0972	2.3393	$S_{\text{TSE},15}$	1.7682	4.4389	3.0406
$S_{\text{NYSE},16}$	2.2664	2.6545	2.0528	$S_{\text{TSE},16}$	1.9065	4.3880	3.0011
$S_{\text{NYSE},17}$	2.2813	2.7824	2.1058	$S_{\text{TSE},17}$	2.1732	4.8737	3.2946
$S_{\text{NYSE},18}$	4.5069	5.5240	5.7066	$S_{\text{TSE},18}$	2.1746	4.7862	3.3036
$S_{\text{NYSE},19}$	1.8315	2.8630	2.2293	$S_{\text{TSE},19}$	2.2255	4.4604	3.1621
$S_{\text{NYSE},20}$	5.3000	5.6623	6.3081	$S_{\text{TSE},20}$	1.8554	4.3601	3.0257
$S_{\text{NYSE},21}$	2.3198	2.7973	2.1474	$S_{\text{TSE},21}$	1.6668	4.2875	2.9623
$S_{\text{NYSE},22}$	2.2613	3.2026	2.3916	$S_{\text{TSE},22}$	2.0788	4.5036	3.1099
$S_{\text{NYSE},23}$	2.7386	2.8809	2.3177	$S_{\text{TSE},23}$	2.0384	4.4373	3.1094
$S_{\text{NYSE},24}$	1.8536	2.9950	2.2968	$S_{\text{TSE},24}$	1.9518	4.1811	2.8882
AVG	2.6229	3.3600	2.8297	*AVG*	2.0026	4.5007	3.1253

Source: MacLean, Foster and Ziemba (2005).

4. Growth strategies

Accurate forecasts for future returns are the foundation for investment strategies. The accumulated wealth process from time t, $W(\tau)$, $\tau \geqslant t$, is defined by the price dynamics and the investment strategy. The decision at time t controls the stochastic dynamic wealth process until a decision is made to revise the strategy. A standard approach in portfolio problems is to set a planning horizon at time T, and specify performance criteria for $W(T)$ as the basis of decision. The decision may be reconsidered between time t and T, either at regular intervals or at random times determined by the trajectory of wealth, but the emphasis is on the horizon. However, the path to the horizon is important, and paths with the same terminal wealth can be very different from a risk/survival

perspective. Conditions which define acceptable paths can be imposed. For example, drawdown constraints are used to avoid paths with a substantial fall-off in accumulated capital (MacLean et al., 2004; Grossman and Zhou, 1993).

4.1. The Kelly strategy

If the distribution of accumulated capital (wealth) at the horizon is the criterion for deciding on an investment strategy, then the rate of growth of capital becomes the determining factor when the horizon is distant. In fact, the optimal growth strategy is the unique *evolutionary stable* strategy, in the sense that it overwhelms other portfolio rules. Strategies that survive in the long run must converge to the optimal growth strategy (Hens and Schenk-Hoppe, 2005). Consider then the average growth rate between t_1 and t_2, for strategy $X = (X(t_1), \ldots, X(t_1 + (n-1)d))$,

$$EG_n(X) = \frac{1}{n}\sum_{i=0}^{n-1} E\ln\bigl(R'\bigl(t_1 + id, t_1 + (i+1)d\bigr)X(t_1 + id)\bigr). \tag{38}$$

The case usually discussed is where the incremental returns are serially independent. So the maximization of $EG_n(X)$ is

$$\max\bigl\{E\ln\bigl(R'\bigl(t_1 + id, t_1 + (i+1)d\bigr)X(t_1 + id)\bigr)\bigr\}, \tag{39}$$

separately for each i. If the returns distribution is the same for each i, a fixed strategy holds over time. The strategy which solves (39) is called the or optimal growth strategy.

In the continuous time case with infinitesimal increments which are defined by the hierarchal Bayes model, the decision is determined from

$$\max\left\{(\phi - re)'\widetilde{X} + r - \frac{1}{2}\widetilde{X}'\Delta\widetilde{X}\right\}, \tag{40}$$

where $\phi_i = \lambda_i + \frac{1}{2}\delta_i^2$, $i = 1, \ldots, K$. Also, $\widetilde{X}' = (x_1, \ldots, x_K)$ is the investment in risky assets. For this continuous time problem, the Kelly strategy is

$$\widetilde{X}^* = \Delta^{-1}(\phi - re). \tag{41}$$

The Kelly or log optimal portfolio is $X^{*\prime} = (x_0^*, \widetilde{X}^{*\prime})$, where $x_0^* = 1 - \sum_{i=1}^K x_i^*$. The continuous time formula can be viewed as an approximate solution to the discrete time investment problem. The Kelly strategy is a *fixed mix*. That is, the fraction of wealth invested in assets is determined by X^*, but rebalancing as wealth varies is required to maintain the fractions.

A variation on the Kelly strategy is the fractional Kelly strategy defined as $\widetilde{X}_f = f\widetilde{X}^*$, $f \geqslant 0$. The fractional Kelly strategy has the same distribution of wealth across risky assets, but varies the fraction of wealth invested in those risky assets. The optimal growth/Kelly strategy has been studied extensively. A summary of its good and bad properties is given in Table 4.

Table 4
Good and bad properties of the optimal growth strategy

Feature	Property	Reference
Good	Maximizes rate of growth	Breiman (1960, 1961); Algeot and Cover (1988).
	Maximizes median log wealth	Ethier (1987).
	Minimizes expected time to asymptotically large goals	Breiman (1961); Algeot and Cover (1988); Browne (1997).
	Never risks ruin	Hakansson and Miller (1975).
	Absolute amount invested is monotone in wealth	MacLean, Ziemba and Li (2005).
	On average never behind any other investor	Finkelstein and Whitley (1981).
	The chance of being ahead of any other investor at least 0.5	Bell and Cover (1980).
	Wealth pulls way ahead of wealth of other strategies	Ziemba and Hausch (1986); MacLean, Ziemba and Blazenko (1992); Amir et al. (2005).
	Kelly is the unique evolutionary strategy	Hens and Schenk-Hoppe (2005).
	Growth optimal policy is myopic	Hakansson (1971), Mossin (1968).
	Can trade growth for security with negative power utility or equivalently with fractional Kelly or risk aversion	Chopra and Ziemba (1993), MacLean and Ziemba (1999); MacLean, Ziemba and Li (2005).
	Optimal fractional Kelly strategy to stay above a growth path with given probability	MacLean et al. (2004); Stutzer (2003).
Bad	It takes a long time to outperform with high probability	Aucamp (1993); Browne (1997), Thorp (2006).
	The investments are extremely large if the risk is low	Ziemba and Hausch (1986)
	The total amount invested swamps the gains	Ethier and Tavare (1983); Griffin (1985); Gottleib (1985).
	There is overinvestment when returns are estimated with error	Chopra and Ziemba (1993); MacLean and Ziemba (1999); Rogers (2001); MacLean, Foster and Ziemba (2002).
	The average return converges to half the return from optimal expected wealth	Ethier and Tavare (1983); Griffin (1985); Gottleib (1985).
	Investing double the Kelly fraction has a zero growth rate	Janacek (1998); Markowitz in Ziemba (2003).
	The chances of losing big in the short term (drawdown) can be high	Ziemba and Hausch (1986).
	Kelly strategy does not optimize expected utility of wealth. Example: Bernoulli trials with $1/2 < p < 1$ and $u(w) = w$. Then $x = 1$ maximizes $u(w)$, but $x = 2p - 1$ maximizes $E \ln(w)$.	Samuelson (1971); Thorp (1971, 1975).

The Kelly portfolio is connected to the fundamental concepts in portfolio theory Luenberger (1993, 1998). If the instantaneous expected return on the Kelly portfolio is μ^*, and the instantaneous expected return on an asset is μ_i, then $\mu_i - r = \beta_i^*(\mu^* - r)$, where $\beta_i^* = \frac{\sigma_{i,*}}{\sigma^{*2}}$. In that expression, $\sigma_{i,*}$ is the covariance between the ith asset and

the Kelly portfolio, and σ^{*^2} is the variance of the Kelly portfolio. This is exactly the CAPM equation, with the Kelly in the role of the market portfolio. If the asset is a derivative with price $y_i = F(P_i, t)$, with P_i the price of asset i, then the CAPM equation implies $\mu_{y_i} - r = \beta^*_{y_i}(\mu^* - r)$. With some rearrangement, this equation becomes $\frac{\partial F}{\partial t} + \frac{\partial F}{\partial P_i} r P_i + \frac{1}{2}\frac{\partial^2 F}{\partial^2 P_i}\sigma^{*^2} P_i^2 = rF$, which is the Black–Scholes pricing formula for derivatives.

The Kelly strategy is central to the theory of capital growth, particularly in the long run. However, it can result in wealth trajectories which are unfavorable (Stutzer, 2004). As a measure of *underperformance*, consider the probability that wealth does not exceed a level defined by a fixed rate of return ρ. Whenever the growth rate $\frac{1}{T}\sum_{t=1}^{T} \ln E[R'(t)X(t)]$ exceeds the target growth rate $\ln \rho$, this underperformance eventually decays to zero exponentially fast as $T \to \infty$. The exponential decay rate of this probability is

$$D(X, \rho) = \max_{\theta>0}\left\{\lim_{T\to\infty}\left\{-\frac{1}{T}\ln E\big(\phi_T(\theta)\big)\right\}\right\}, \tag{42}$$

where

$$\phi_T(\theta) = \exp\left(-\theta\sum_{t=1}^{T}\big(\ln R(t)'X(t) - \ln \rho\big)\right). \tag{43}$$

The strategy $X_{\max}$ that maximizes $D(X, \rho)$ will result in the lowest large-T probabilities of failing to exceed the constantly growing wealth target $W(0)\rho^T$. When the $R(t)$ are independent and identically distributed, the solution $X_{\max}$ can be found from the problem

$$\max_{\theta>0}\left\{-\frac{1+\theta}{\rho}\max_X E\left[\frac{1}{1+\theta}(R'X)^{-\theta}\right]\right\}. \tag{44}$$

The $\theta_{\max}$ is an endogenously determined parameter in a power utility function.

In terms of the probability of not exceeding a target growth rate ρ, the Kelly strategy might not be preferable (Stutzer, 2003). In an example with ($\mu = 0.15$, $\sigma = 0.3$) for a lognormal asset and a risk free asset with $r = 0.07$, Stutzer compared the Kelly strategy which invests 89% of capital in stock with portfolios with 47% of capital in stock and with 33% in stock; see Figure 3. The target growth rate was set at $\ln(\rho) = 0.08$.

In this example the Kelly is dominated by a fractional Kelly strategy with 47% in stock ($f = 0.528$) at each planning horizon from a risk perspective. The investment fraction is linked to risk or the underperformance probability.

The concept of concentrating on wealth above a target path is also referred to as *the discretionary wealth hypothesis* (Wilcox, 2005). The investment objective is to maximize the utility of wealth in excess of the amount that would trigger a shortfall. The discretionary wealth model would have a log wealth objective and a constraint requiring wealth to exceed the shortfall level. Variations on this model, and the connection to fractional Kelly strategies are explored later in this chapter.

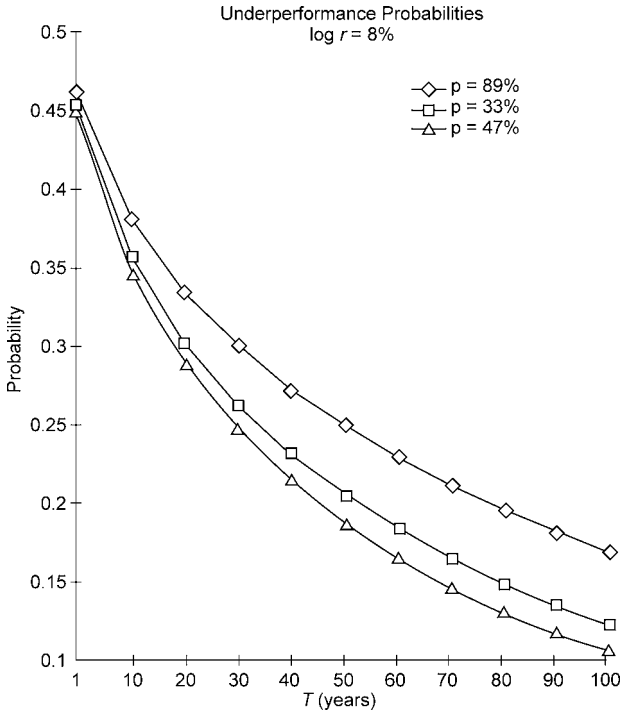


Fig. 3. Kelly underperformance probability. Source: Stutzer (2003).

A general framework for simultaneously considering preferences for shortfall risk and return is stochastic dominance, which is discussed in the next section. That material is presented for the case where the distribution of returns is known. In many situations, the risks associated with the Kelly strategy are too high for known distributions. If the estimation of returns contains significant error, the motivation for a more conservative strategy such as fractional Kelly is even stronger. To see this, recall that the risk of an estimator is given as $R(\lambda, \hat{\lambda}) = \operatorname{tr} E(\hat{\lambda} - \lambda)(\hat{\lambda} - \lambda)'$. If the optimal estimator is λ^*, then $R(\lambda, \hat{\lambda}) = R(\lambda, \lambda^*) + R(\hat{\lambda})$, with $R(\hat{\lambda})$ being the additional risk for $\hat{\lambda}$. Consider a contamination of the estimator $\hat{\lambda}$ resulting from adding error. The additional risk that is incurred by using a contaminated estimate $\hat{\lambda}_{\text{cont}}$ can be represented by the relative savings loss for the estimator $\hat{\lambda}_{\text{cont}}$

$$RSL\big(\hat{\lambda}_{\text{cont}}\big) = \frac{R(\hat{\lambda})}{R(\hat{\lambda}_{\text{cont}})}. \tag{45}$$

Consider an investor with wealth w_t at time t and as an investment strategy using the fractional optimal growth: $\widetilde{X}_f = f\widetilde{X}^*$. Let the $(t+1)$th period wealth be $W_{\hat{\lambda}}(t+1)$ and $W_{\lambda^*}(t+1)$ for the growth strategy with the estimated rate of return $\hat{\lambda}$ and the optimal

estimate λ^*, respectively. The ratio

$$WL(X_{\hat{\lambda}}) = E \log \frac{W_{\hat{\lambda}}(t+1)}{W_{\lambda^*}(t+1)} \tag{46}$$

indicates the wealth loss from the estimate $\hat{\lambda}$ for the rates of return.

With $\widetilde{X}_{\hat{\lambda}}(t)$, $\widetilde{X}_{\hat{\lambda}_{\text{cont}}}(t)$ being the fractional Kelly strategies from using standard and contaminated estimates for the rate of return, the one period ahead relative wealth loss from a contaminated estimate is defined as

$$RWL(X_{\hat{\lambda}_{\text{cont}}}) = \frac{WL(X_{\hat{\lambda}})}{WL(X_{\hat{\lambda}_{\text{cont}}})}. \tag{47}$$

With greater contamination (poorer estimate of mean rate of return), the relative savings loss and the relative wealth loss both increase. However, for the growth strategies the wealth loss is modulated by the risk aversion index and thus the fraction in the Kelly strategy. That is, for a fixed contamination

$$\begin{aligned} RWL(X_{\hat{\lambda}_{\text{cont}}}) &< RSL\big(\hat{\lambda}_{\text{cont}}\big), \quad f < 1, \\ RWL(X_{\hat{\lambda}_{\text{cont}}}) &= \text{RSL}\big(\hat{\lambda}_{\text{cont}}\big), \quad f = 1, \\ RWL(X_{\hat{\lambda}_{\text{cont}}}) &> RSL\big(\hat{\lambda}_{\text{cont}}\big), \quad 1 > f > 1. \end{aligned}$$

For proof of this result see MacLean, Foster and Ziemba (2005). The wealth loss depends upon the risk aversion at the time of decision. In the decision rule, the risk aversion parameter β defines a fraction of capital invested in the optimal growth portfolio. When $f < 1$, the control of decision risk also reduces the impact of estimation error. Correspondingly, when $f > 1$, the overinvestment increases the effect of estimation error. The risk reduction effects of fractional Kelly strategies are separate from the reduction in estimation error through improved estimation methodologies.

4.2. Stochastic dominance

Usually an investors' preferences for wealth over time are the basis for the investment decision. Those preferences involve growth of capital, but also an aversion to serious decay in wealth. The traditional method for defining preferences for wealth at the planning horizon is expected utility. The uncertain accumulated capital at the horizon with various strategies can be compared using that criterion. Wealth $W_1(T)$ dominates $W_2(T)$ iff $Eu(W_1(T))$ is greater than or equal to $Eu(W_2(T))$ for every $u \in U$, with strict inequality for at least one u (Hanoch and Levy, 1969).

There are alternative orders of dominance based on the class of utility function. The classes are defined by the sign of the kth order derivative of the utility function. Consider

$$U_k = \big\{u \mid (-1)^{j-1} u^{(j)} \geqslant 0,\ j = 1, \ldots, k\big\}. \tag{48}$$

So U_1 is the class of monotone nondecreasing functions, U_2 is the class of concave, monotone nondecreasing functions, and so on, with $U_1 \supseteq U_2 \supseteq \cdots$. If $U_\infty =$

$\lim_{j\to\infty} U_j$, then U_∞ contains the set of power utility functions, $U_* = \{u \mid u(w) = c_1 w^{c_2}, c_1 > 0, c_2 < 1\}$. This special class is important since it captures features such as risk aversion, and positive skewness, and is amenable to analytic results. The utility $u(w) = \ln(w)$ is a member of the class of power utilities, and is the utility generating the Kelly investment strategy. From a utility perspective it is quite specialized, although it generates the most systematic and dominant portfolio.

Increasingly restrictive orders of dominance are defined by the classes of utilities. Consider the dominance criteria:

Wealth $W_1(T)$ kth order dominates wealth $W_2(T)$, denoted $W_1(T) \succeq_k W_2(T)$, if $Eu(W_1(T)) \geqslant Eu(W_2(T))$ for all $u \in U_k$, with strict inequality for at least one u.

Another formulation for stochastic dominance follows from the distribution functions for wealth. Let $W_1(T)$ and $W_2(T)$ have densities f_1 and f_2, respectively. For a density f, consider for $k = 1, 2, \ldots$ the nested integrations

$$I_k^{\mathsf{f}}(w) = \int_{-\infty}^{w} I_{k-1}^{\mathsf{f}}(z)\,\mathrm{d}z, \tag{49}$$

where $I_0 = \mathsf{f}$. Then

Wealth W_1 kth order dominates wealth W_2 iff $I_k^{\mathsf{f}_1}(w) \leqslant I^{\mathsf{f}_2}(w)$, for all w, with strict inequality for at least one w.

It is convenient to write the nested integral in terms of loss functions. Again consider wealth W, with distribution F. Let $\rho_1(\alpha) = \inf\{\eta \mid F(\eta) \geqslant \alpha\}$, the αth percentile of the wealth distribution. For $k \geqslant 2$, define

$$\rho_k^W(\alpha) = \int_{-\infty}^{\rho_{k-1}(\alpha)} \big(\rho_{k-1}(\alpha) - w\big)^{k-1}\mathrm{d}F(w), \tag{50}$$

which is a normed measure of the loss along the distribution. This loss function is a rearrangement of the nested integration. Then

Wealth W_1 kth order dominates wealth W_2 iff for $k \geqslant 2$, $\rho_k^{W_1}(\alpha) \leqslant \rho_k^{W_2}(\alpha)$, for all α, with strict inequality for at least one α. For $k = 1$, $\rho_1^{W_1}(\alpha) \geqslant \rho_1^{W_2}(\alpha)$, for all α, with strict inequality for at least one α.

The variations on the dominance relation provide insight into the conditions for the wealth process. The loss function definition of dominance is particularly useful since all the measures have the same argument α, and values of α are easy to comprehend. It is clear that $\rho_1^W(\alpha)$, $0 \leqslant \alpha \leqslant 1$, is the inverse cumulative distribution, and ordering based on this measure expresses a preference for wealth. Since $\int_{-\infty}^{\rho_1(\alpha)} (\rho_1(\alpha) - w)\,\mathrm{d}F(w) = \alpha\rho_1(\alpha) - \int_{-\infty}^{\rho_1(\alpha)} w\,\mathrm{d}F(w)$, the second order measure,

$\rho_2^W(\alpha), 0 \leqslant \alpha \leqslant 1$, is equivalent to the Lorenz curve (Lorenz, 1905). It is defined as $L(\alpha) = \int_{-\infty}^{\rho_1(\alpha)} w \, \mathrm{d}F(w)$, and reflects risk aversion in that it is sensitive to by the lower tail of the distribution. The measure $\rho_3^W(\alpha), 0 \leqslant \alpha \leqslant 1$, captures the aversion to variance. These characteristics are components of the preferences which investors express with a utility function.

4.3. Bi-criteria problems: Fractional Kelly strategies

In general, the ordering of wealth distributions at the horizon using stochastic dominance is not practical. Rather than use the full range of α values with the moments definition, it is more realistic to identify specific values and work with a vector order. The vector ordering $(x_1, \ldots, x_p) < (y_1, \ldots, y_p)$ holds iff $x_i \leqslant y_i$, $i = 1, \ldots, p$, with at least one strict inequality. Selecting two values of α to set up bi-criteria problems $(\rho_k(\alpha_1), \rho_k(\alpha_2))$, has some appeal. Consider terminal wealth variables $W_1(T)$ and $W_2(T)$, and values $\alpha_1 < \alpha_2$.

For $k \geqslant 2$, $W_1(T)$ is kth order (α_1, α_2)-preferred to $W_2(T)$, denoted $W_1(T) \gg_k W_2(T)$, iff $(\rho_k^{W_1}(\alpha_1), \rho_k^{W_1}(\alpha_2)) < (\rho_k^{W_2}(\alpha_1), \rho_k^{W_2}(\alpha_2))$. $W_1(T)$ is 1st order (α_1, α_2)-preferred to $W_2(T)$ iff $(\rho_1^{W_1}(\alpha_1), \rho_1^{W_1}(\alpha_2)) > (\rho_1^{W_2}(\alpha_1), \rho_1^{W_2}(\alpha_2))$.

The preference ordering is *isotonic* with respect to the dominance ordering: $W_1(T) \succeq_k W_2(T) \Rightarrow W_1(T) \gg_k W_2(T)$ (Orgyczak and Ruszczynski, 2002).

The preference ordering on wealth leads to analogous optimization problems for determining a preferred investment strategy. The definition of a problem requires the specification of lower and upper α-values, α_L and α_U, capturing risk and return. There are natural values based on conventional use. So $\rho_1(\alpha_U) = \rho_1(0.5) =$ median wealth, which is equivalent to the mean log-wealth $= E(\ln(W(T)))$. If the lower value is α_L, then $\rho_1(\alpha_L) =$ the $100\alpha_L$th percentile. A bi-criteria problem defined by the first order preference ordering is P_{VaR}:

$$\mathrm{Max}\big\{E\big(\ln\big(W(T)\big)\big) \mid \Pr\big[W(T) \leqslant w_{\mathrm{VaR}}\big] \leqslant \alpha\big\}. \tag{51}$$

This is a variation on the unconstrained optimal growth problem, with a value-at-risk constraint imposed. Since $\{W(T) \leqslant w_{\mathrm{VaR}}\} \Leftrightarrow \{\ln(W(T)) \leqslant \ln(w_{\mathrm{VaR}})\}$, the constraint is analogous to a condition on the growth rate of trajectories as in the underperformance measure (Stutzer, 2003).

For second order preference, let $\rho_2(\alpha_U) = \rho_2(1.0) = E(W(T)) =$ mean wealth. With $\rho_2(\alpha_L) =$ the lower 100αth percent incomplete mean, the second order preference problem is P_{CVaR}:

$$\mathrm{Max}\big\{E\big(W(T)\big) \mid E\big(W(T)I_{\alpha_L}\big) \geqslant w_{\mathrm{CVaR}}\big\}. \tag{52}$$

In this notation I_{α_L} is an indicator variable for the lower $100\alpha_L\%$ of the wealth distribution. The CVaR measure is developed by Rockafellar and Uryasev (2000). This

Table 5
Path measures

Criterion	Specification	Path measure
Change	$b < 0$	$\delta_1(b) = \Pr[\frac{1}{W(\tau)}\, \mathrm{d}W(t+\xi) > b\xi, \xi > 0]$
Levels	(w_L, w_U)	$\delta_2(w_L, w_U) = \Pr[\tau_{w_U} < \tau_{w_L}]$

Source: MacLean, Zhao and Ziemba (2006).

Table 6
Alternative decision models

Model	Criterion	Problem
M_1	Expected utility:power utility with risk aversion index $\frac{1}{1-\beta}$, $\beta < 1$.	$\mathrm{Max} E[\frac{1}{1-\beta} W(T)^{\beta}]$
M_2	First order dominance: optimal median wealth subject to a VaR constraint.	$\mathrm{Max}\{E(\ln(W(T))) \mid \Pr[W(T) \leqslant w_{\mathrm{VaR}}] \leqslant \alpha_L\}$
M_3	Second order dominance: optimal mean wealth subject to a CVaR constraint.	$\mathrm{Min}\{E(W(T)) \mid E(W(T) I_{\alpha_L}) \geqslant w_{\mathrm{CVaR}}\}$
M_4	Drawdown: optimal median wealth subject to a drawdown constraint.	$\mathrm{Max}\{E(\ln(W(T))) \mid \delta_1(b) \geqslant 1 - \alpha_L\}$
M_5	Wealth goals: optimal median wealth subject to control limits.	$\mathrm{Max}\{E(\ln(W(T))) \mid \delta_2(w_L, w_U) \geqslant 1 - \alpha_L\}$

Source: MacLean, Zhao and Ziemba (2006).

problem is discussed in De Giorgio (2005), where the consistency of the objective and constraint combination for second order dominance is established.

The preference problems are defined by wealth at a horizon time. The properties of wealth at the horizon are significant, but the path of wealth to the horizon is also important as was indicated earlier with the discretionary wealth hypothesis. There are unsustainable wealth levels and losses, so measures which reflect the chance of survival should be included in the decision process. There are two approaches considered: (i) acceptable paths based on the *rate of change* or growth in wealth; (ii) acceptable paths based on wealth *levels*. The chance that a path is acceptable defines a path measure. Table 5 gives a statement of such measures. The notation τ_w refers to the first passage time to the wealth level w.

The change measure puts a lower bound on the fallback in any period. The levels measure requires trajectories to reach an upper level before falling to a lower level, e.g., doubling before halving. These measures could be used in place of a moment measure or even in addition to those measures to control the path of wealth. A variety of decision models for determining an investment strategy, which have a form of optimality referred to as growth-security efficiency, are displayed in Table 6.

The usual emphasis in the bi-criteria problems is on median wealth. That is important since the unconstrained solution to a median wealth problem is the Kelly strategy. The problems in Table 6 are defined for continuous or discrete time models. Information on prices is recorded at discrete points in time, and decisions on allocation of capital to investment instruments are taken at discrete points in time. However, wealth accumulates continuously and it reasonable to analyze capital growth in continuous time. So discrete price data is used to estimate parameters in a continuous time pricing model, but the forecast growth in capital between decision points is continuous. There are some limitations to this continuous time parametric approach. It is possible that there is not a good model for prices, and discrete time and state scenarios are better able to capture future prospects. Also, certain pseudo-investment instruments such as lotteries and games of chance are discrete by nature. In the analysis of investment strategies in this section, it is assumed that the continuous time Bayesian pricing model is correct, although knowledge of the model parameters is uncertain. At the end of this section an example of a discrete scenario, discrete time problem is considered.

For continuous time investment with asset prices defined by geometric Brownian motion, the class of feasible investment strategies for the single risk-free and m risky opportunities is

$$\chi_t = \left\{ X(t) \mid \left[X(t)'\big(\hat{\phi}(t) - re\big) + r - \frac{1}{2} X(t)' \hat{\Delta} X(t) \right] \geqslant 0 \right\}. \tag{53}$$

The condition in (53) is for positive growth. The expected growth rate for wealth, given an investment strategy $X(t) \in \chi_t$, is defined as $G(X) = E \ln W(T)^{1/T}$. The Kelly or optimal growth strategy is defined by $X^*(t) = \arg\max G(X)$. In the continuous time problem, the Kelly strategy, defining the investments in risky assets, has the closed form $X^*(t) = \Delta^{-1}(\hat{\phi}(t) - re)$, where $e' = (1, \ldots, 1)$. The subclass of fractional Kelly strategies is defined as

$$\chi_t^* = \big\{ X(t) \mid X(t) = f X^*(t),\ f \geqslant 0 \big\}. \tag{54}$$

Since $X^*(t) \in \chi_t$, then $\chi_t^* \subseteq \chi_t$. The significance of the fractional Kelly strategies lies in their optimality for the problems in Table 6, assuming the Bayesian geometric Brownian motion model for prices is correct.

Let $X_{M_j}(t)$ be the optimal solution to growth problem M_j, $j = 1, \ldots, 5$, defined in Table 6. Then $X_{M_j}(t) \in \chi_t^*$, that is, the solution is fractional Kelly.

For proof see MacLean, Zhao and Ziemba (2005).

In the continuous time formulation, the optimal investment strategies for the various problems have the same form. However, the actual fraction in each problem, which controls the allocation of capital to risky and risk-free instruments, depends on the decision model and parameters. The formulas for the fractions for different models are displayed in Table 7. The notation $\tilde{\mu} = (\hat{\phi} - re)' X^*(t) + r$, and $\tilde{\sigma}^2 = X^{*'}(t) \Delta X^*(t)$ is used for the

Table 7
Investment fractions

Model	Parameters	Fraction
M_1	β	$f_1 = \frac{1}{1-\beta}$
M_2	(ρ_1^*, α)	$f_2 = \frac{B_1+\sqrt{B_1^2+2A_1C_1}}{A_1}$ $A_1 = (\hat{\phi} - re)' \Delta^{-1} (\hat{\phi} - re)$ $B_1 = A_1 + z_\alpha \sqrt{\frac{A_1}{T-t}}$ $C_1 = r - (T-t)^{-1} \ln \frac{\rho_1^*}{w_t}$
M_3	(ρ_2^*, α)	$f_3 = \frac{B_2+\sqrt{B_2^2+2A_2C_2}}{A_2}$ $A_2 = \left[z_\alpha \frac{\Phi'(z_\alpha)}{\alpha} + \left(\frac{\Phi'(z_\alpha)}{\alpha}\right)^2\right] A_1$ $B_2 = A_1 - \frac{\Phi'(z_\alpha)}{\alpha} \sqrt{\frac{A_1}{T-t}}$ $C_2 = r - (T-t)^{-1} \ln \frac{\rho_2^{**}}{w_t}$
M_4	(b, α)	$f_4 = \frac{b}{\tilde{\mu} + z_\alpha \tilde{\sigma}}$
M_5	(w_L, w_U)	$f_5 = h_t \cdot \widetilde{H} + \sqrt{[h_t \cdot \widetilde{H}]^2 + \frac{2rh_t}{\tilde{\sigma}^2}}$ $\widetilde{H} = \frac{\tilde{\mu}-r}{\tilde{\sigma}^2}$ $h_t = \frac{\ln(w_t)-\ln(w_L)}{\ln(w_t)-\ln(y^* w_L)}$

Source: MacLean, Zhao and Ziemba (2006).

mean and variance of the rate of return on the Kelly strategy. Also y^* is the minimum positive root of the equation $\gamma y^{c+1} - y + (1-\gamma) = 0$, for $c = \frac{\ln(w_U)-\ln(w_t)}{\ln(w_t)-\ln(w_L)}$.

Although it is not obvious from the formulas, the fractions in Table 7 reflect the orderings from the sequence of stochastic dominance relations. So the fraction $f_1 = \frac{1}{1-\beta}$ in M_1 is the most specific, but has fewer degrees of freedom in defining risk. There is a set of specifications for (ρ_1^*, α) in M_2, for example, which yield the same fraction, that is $f_2 = f_1$. In general, if the settings for model M_i are represented by π_i, there are equivalence classes of settings

$$\Pi_i(\beta) = \left\{ \pi_i \mid \arg\max\big(M_i(\pi_i)\big) = \frac{1}{1-\beta} X^* \right\}, \quad i = 2, \ldots, 5. \tag{55}$$

The fractional Kelly strategy arises from aversion to risk, which can be characterized by a power utility or by explicit risk constraints. Figure 3 showed that the Kelly strategy was dominated from an underperformance probability perspective. The definition of the underperformance problem with iid returns in Equation (44) leads to the result that the strategy which minimizes the underperformance probability for a target growth rate ρ less than the optimal growth rate, is fractional Kelly. Alternatively, the strategy which maximizes the probability of outperforming a target growth rate is fractional Kelly. The wealth beyond the target growth path has been termed *discretionary*

Table 8
Daily rates of return

	Stocks	Bonds	Cash
Mean	0.00050	0.00031	0.00019
Variance	0.00062	0.00035	0.000
Covariance	0.000046		

Source: MacLean, Zhao and Ziemba (2005).

wealth (Wilcox, 2003). So the fractional Kelly strategy maximizes the probability of discretionary wealth, a result referred to as the *discretionary wealth hypothesis*.

4.3.1. Example: Investment in stocks and bonds in continuous time

To illustrate the connection between the investment fraction in the optimal growth strategy and the specifications for ρ_1 and α_L, the allocation of investment capital to stocks, bonds, and cash over time is considered. Hensel, Ezra and Ilkiw (1991) and Blake, Lehmann and Timmermann (1999) show that it is strategic asset allocation that is the dynamic rebalancing of stocks, bonds and cash that determine most of the mean and variance of portfolio returns rather than individual stock selection. Daily prices were generated for 260 trading days, based on the statistics in Table 8. They were determined from statistics on total returns from the S&P500, Solomon Brothers bond index, and US treasury bills.

Initial wealth was set at \$1. For a trajectory of simulated prices, an investment strategy was determined every 10 days. At a rebalance time, the data on past prices was used to update estimates for model parameters based on the empirical Bayes methodology. With revised estimates, the continuous time formulas were used to calculate the investment strategy. The fractions of wealth invested in the optimal portfolio of stocks and bonds in the initial period is shown in Table 9. The fractions are particularly sensitive to the VaR level ρ_1. The strategies are fractional Kelly, since the distribution of investment capital in (stocks, bonds) is (0.6, 0.4) for each combination of (ρ_1, α_L).

With 1000 trajectories simulated, the average wealth at the end of the planning horizon was calculated; see Table 10.

This example is intended for illustration of the effect of risk constraints such as VaR on the fractional Kelly strategy and the accumulated capital. The impact of the VaR level, ρ_1^* is most pronounced. Strict requirements in the VaR level lead to very conservative fractional Kelly strategies.

4.3.2. Example: Investment in stocks and bonds in discrete time

For comparison with the continuous time problem, M_2 was implemented in discrete time, with years as time units. The statistics on annual returns from the data are in Table 11.

Table 9
VaR strategies (stocks, bonds)

ρ_1^*	α_L				
	0.01	0.02	0.03	0.04	0.05
0.99	(0.30, 0.20)	(0.34, 0.23)	(0.38, 0.25)	(0.40, 0.27)	(0.43, 0.29)
0.98	(0.56, 0.37)	(0.64, 0.42)	(0.69, 0.46)	(0.74, 0.50)	(0.79, 0.53)
0.97	(0.82, 0.54)	(0.93, 0.62)	(10.01, 0.68)	(10.09, 0.72)	(10.16, 0.77)
0.96	(1.08, 0.72)	(1.22, 0.81)	(1.33, 0.89)	(1.43, 0.95)	(1.52, 1.01)
0.95	(1.34, 0.89)	(1.52, 1.01)	(1.65, 1.10)	(1.78, 1.18)	(1.89, 1.26)

Source: MacLean, Zhao and Ziemba (2005).

Table 10
Expected wealth at the horizon

ρ_1^*	α_L				
	0.01	0.02	0.03	0.04	0.05
0.99	1.0629	1.0649	1.0666	1.0681	1.0696
0.98	1.0764	1.0797	1.0826	1.0854	1.0884
0.97	1.0881	1.0953	1.1019	1.1084	1.1154
0.96	1.1053	1.1189	1.1310	1.1432	1.1557
0.95	1.1292	1.1509	1.1703	1.1895	1.2091

Source: MacLean, Zhao and Ziemba (2005).

Table 11
Annual return statistics

Parameter	Stocks	Bonds	Cash
Mean	0.08750	0.0375	0
Variance	0.1236	0.0597	0
Correlation	0.32		

Source: MacLean, Zhao and Ziemba (2005).

A corresponding set of scenarios was created (sampling from a lognormal distribution for stocks and bonds), and they are displayed in Table 12. The sampling process was structured so that sample statistics were as close as possible to the statistics in Table 10 (MacLean et al., 2004).

The planning horizon was 3 years and the same scenarios were used each year. There were 64 scenarios, each with a probability of $1/64$. With this discrete time and discrete

Table 12
Return scenarios

Stocks	Bonds	Cash	Probability
0.95	1.015	1	0.25
1.065	1.100	1	0.25
1.085	0.965	1	0.25
1.250	1.070	1	0.25

Source: MacLean, Zhao and Ziemba (2005).

Table 13
Investment strategy and rate of return

ρ_{1*}	Year 1			Year 2			Year 3			Yearly
	Stocks	Bonds	Cash	Stocks	Bonds	Cash	Stocks	Bonds	Cash	return
0.950	1	0	0	0.492	0.508	0	0.492	0.508	0	1.061
0.970	1	0	0	0.333	0.667	0	0.333	0.667	0	1.057
0.990	0.456	0.544	0	0.270	0.730	0	0.270	0.730	0	1.041
0.995	0.270	0.730	0	0.219	0.590	0.191	0.218	0.590	0.192	1.041
0.999	0.270	0.730	0	0.008	0.020	0.972	0.008	0.020	0.972	1.017

Source: MacLean et al. (2004).

scenario formulation, problem M_2 was solved with $\alpha = 0.01$ and a variety of values for the VaR level ρ_1^*. Starting wealth was \$1. The results from solving the problems are shown in Table 13. (Details on this problem are in MacLean et al., 2004.) If the annual returns are compared to the results for the continuous time problem with the same α, the returns in this case are slightly lower. However, the continuous time application involved rebalancing every 10 days. In Table 13, it can be observed that the very strict VaR condition almost eliminates the possibility of growth. As well, the optimal strategy is not fractional, with the investment mix changing as the horizon approaches.

4.4. *Growth-security trade-off*

The solutions displayed in Table 7 are derived from the continuous time wealth equation, although the strategies are calculated at discrete decision points in time. The alternative problems in Table 6 can be based on the discrete time wealth equation, but the optimal solution is not necessarily fractional Kelly. That point was demonstrated in the discrete time example. However, the fractional Kelly solution may be near-optimal. If the feasible strategies for the discrete time problem are restricted to the class of fractional strategies, the solutions are *effective* (MacLean, Ziemba and Blazenko, 1992; MacLean and Ziemba (1991, 2000)). That is, as the fraction changes, the growth (objec-

Table 14
Computation of measures

Measure	Specification	Formula
$\rho_1(\alpha_U)$	$\alpha_{U=0.5}$	$w_t \exp\{(\mu(X)-r)\lambda(X)+r-\frac{1}{2}X'\Delta XT\}$
$\rho_2(\alpha_U)$	$\alpha_{U=1.0}$	$w_t \exp\{(\mu(X)-r)(\lambda(X)+r)T\}$
$\rho_1(\alpha_L)$	w_{VaR}	$\Phi\big(\frac{D(X)T-\ln\frac{w_{\text{VaR}}}{w_t}}{\sigma(X)\lambda(X)T^{1/2}}\big)$
$\rho_2(\alpha_L)$	w_{CVaR}	$w_t \exp\{(\mu(X)-r)(\lambda(X)+r)T\}[\Phi(z_{\alpha_L})-\sigma(X)]$
$\delta_1(\alpha_L)$	β	$\big[1-\Phi(\frac{\ln\beta-\mu(X)}{\sigma(X)/T^{1/2}})\big]^{nT}$
$\delta_2(\alpha_L)$	(w_L, w_U)	$\frac{1-(\frac{w_U}{w_t})^{2D(X)/\sigma^2(X)\lambda^2(X)}}{1-(\frac{w_U}{w_L})^{2D(X)/\sigma^2(X)\lambda^2(X)}}$

Source: Dohi et al. (1995) and MacLean and Ziemba (1999).

tive) and security (constraint) move in opposite directions, so that growth is monotone non-increasing in security. Specifically, for $0 \leqslant f \leqslant 1$, $\alpha_L < 0.5$, $\alpha_U \geqslant 0.5$,

$$\frac{\mathrm{d}}{\mathrm{d}f}\rho_i(\alpha_L) \geqslant 0, \quad \frac{\mathrm{d}}{\mathrm{d}f}\delta_i(\alpha_L) \geqslant 0, \quad \frac{\mathrm{d}}{\mathrm{d}f}\rho_i(\alpha_U) \leqslant 0, \quad i = 1, 2. \tag{56}$$

The implication of this monotonicity is that growth can be traded for security using the fraction allocated to the optimal growth portfolio. With the fractional Kelly strategy fX^*, the computation of the growth and security measures as a function of f is accomplished with known formulas (MacLean, Ziemba and Blazenko, 1992; Dohi et al., 1995). The approximating continuous time formulas are in Table 14.

So the growth–security trade-off can be observed for various fractional Kelly strategies and suitable fractions (meeting investor preferences) can be determined. This will be demonstrated with some examples.

4.4.1. Example: Investing in turn-of-the-year effect

The trade-off between growth and security was used by Clark and Ziemba (1987) in the analysis of investment in the turn of the year effect. The excess return of small cap stocks minus large cap stocks is most pronounced in January. The distribution of gains at the turn of the year from holding long positions in the Value Line Index of small cap stocks, and short positions in the Standard and Poors Index of large cap stocks is given in Table 15, where each point was worth $500 at that time. The data covers the period 1976–77 to 1986–87.

The Kelly strategy calculated from this distribution invests 74% of ones fortune in the trade. This is very aggressive considering the possible estimation errors and the market volatility.

Table 15
Returns distribution for VL/S&P spread

Points spread	7	6	5	4	3	2	1	0	−1
Probability	0.007	0.024	0.07	0.146	0.217	0.229	0.171	0.091	0.045

Source: Clark and Ziemba (1987).

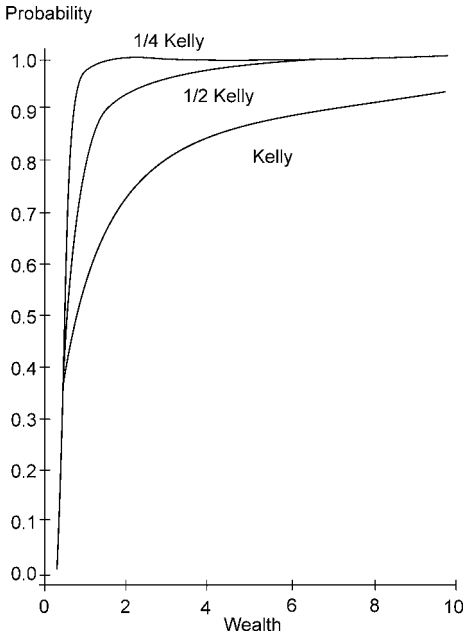


Fig. 4. Turn-of-the-year effect: probability of reaching $10 million before ruin for Kelly, half Kelly and quarter Kelly strategies. Source: MacLean, Ziemba and Blazenko (1992).

In Figure 4 is a graph for fractional Kelly strategies with the turn of the year trade, showing the chance of reaching a wealth of 10 million before ruin, starting from various initial wealth levels. The graph for 0.25 Kelly is much more secure. Similarly in Figure 5, the trade-off between relative growth and probability of achieving a desired wealth level is displayed. Going from Kelly to quarter Kelly realizes a probability gain (security) of about 0.25 and an almost equivalent loss in relative growth. An approximate 0.25 Kelly strategy was used with consistent success in actual trades on this commodity by W.T. Ziemba in the 14 years from 1982/83 to 1996/97, winning each year; see Clark and Ziemba (1987), Ziemba (1994), Hensel and Ziemba (2000) and

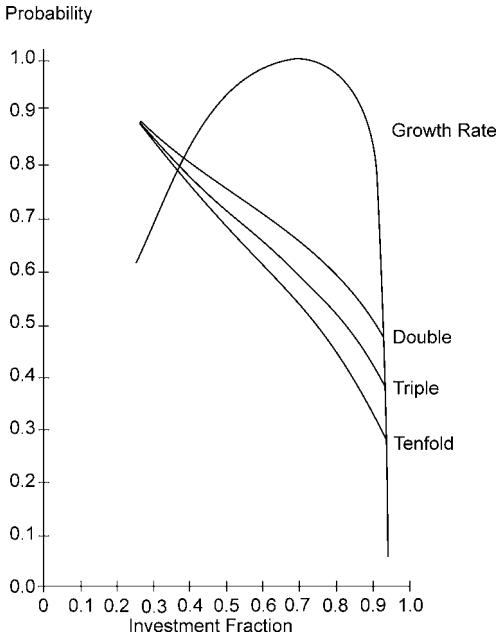


Fig. 5. Relative growth versus probability of wealth. Source: MacLean, Ziemba and Blazenko (1992).

Rendon and Ziemba (2005). Because the declining Value Line volume makes the trade risky, he has not done the trade since 1997.

The current markets have become much more dangerous than during the period of the study of the turn of the year effect: see Rendon and Ziemba (2005) for an update to 2004/5. The January effect still exists, however, in the futures markets, but now is confined to the second half of December.

4.4.2. Example: Kentucky Derby

Donald Hausch and William Ziemba have written extensively on the advantages of the capital growth approach to wagering on horseraces; see Hausch, Ziemba and Rubinstein (1981), Hausch and Ziemba (1985, 1990a, 1990b), Hausch, Lo and Ziemba (1994) and Hausch and Ziemba (1995). They have developed a betting approach, the Dr. Z system, based on the Kelly strategy (Ziemba and Hausch, 1984, 1986, 1987). Figure 6 shows wealth level histories from place and show betting on the Kentucky Derby from 1934 to 1994 using the Dr. Z system (Bain, Hausch and Ziemba, 2006). The system uses probabilities from the simpler win market to determine bets in the more complex place/show market, where inefficiencies are more likely to occur. Starting with initial wealth of $2500, and a 4.00 dosage index filter rule, Kelly and half Kelly strategies are compared

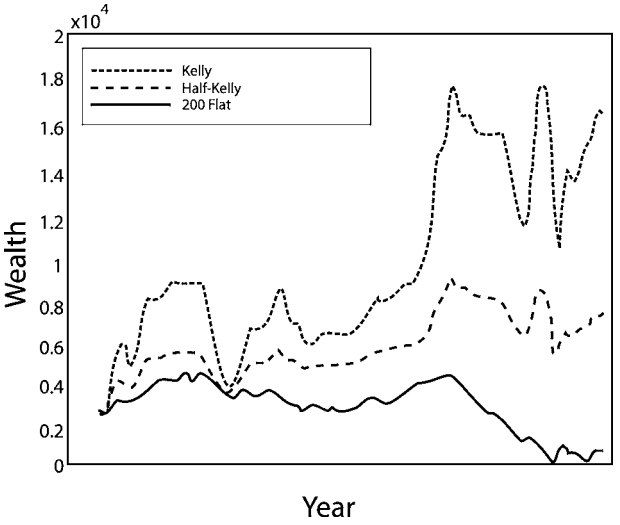


Fig. 6. Kentucky Derby capital growth. Source: Bain, Hausch and Ziemba (2006).

with \$200 flat bets on the favorite. The full Kelly yields a final wealth of \$16,861, while half Kelly has a final wealth of \$6945, but with a much smoother wealth path.

4.4.3. Example: Lotto games

In lotto games players select a small set of numbers from a given list. The prizes are shared by those with the same numbers as those selected in the random drawing. The lottery organization bears no risk in the pari-mutuel system and takes its profits before the prizes are shared. Hausch and Ziemba (1995) survey these games, Ziemba et al. (1986) studied the 6/49 game played in Canada and several other countries. Numbers ending in eight and especially nine and zero tend to be unpopular. Six tuples of unpopular numbers have an edge with expected returns exceeding their cost. See also Chernoff (1980/1981) for this idea in a different type of lotto game. The expected value approaches \$2.25 per dollar wagered when there are carryovers (that is when the Jackpot is accumulating because it has not been won). However, investors may still lose because of mean reversion (the unpopular numbers tend to become less unpopular over time) and gamblers' ruin (the investor has used up his resources before winning). MacLean, Ziemba and Blazenko (1992) investigated how an investor might do playing sets of unpopular numbers with a combined advantage using the data in Table 16.

The optimal Kelly wagers are extremely small. The reason for this is that the bulk of the expected value is from prizes that occur with less than one in a million probability. A wealth level of \$1 million is needed to justify even one \$1 ticket. Figure 7 provides the chance that the investor will double, quadruple or tenfold this fortune before it is

Table 16
Lotto 6/49 data

Prizes	Prob.	Value	Contribution %
Jackpot	1/13983816	$6M	42.9
Bonus	1/2330636	$0.8M	34.3
5/6	1/55492	M	9.0
4/6	1/1032	$5000	14.5
3/6	1/57	$150	17.6
Edge			18.1
Kelly bet			0.00000011
Number of tickets with 10M bankroll			11

Source: MacLean, Ziemba and Blazenko (1992).

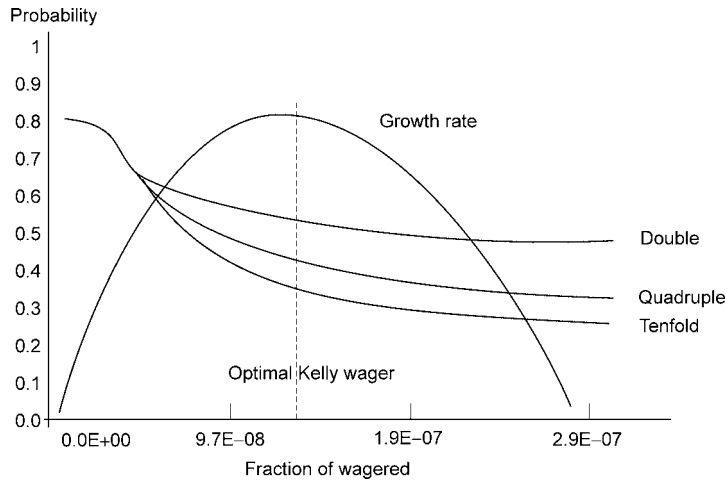


Fig. 7. Lotto 6/49—Probability of multiplying before losing half of ones fortune vs bet size. Source: MacLean, Ziemba and Blazenko (1992).

halved using Kelly and fractional Kelly strategies. These chances are in the 40–60%. With fractional Kelly strategies in the range of 0.00000004 and 0.00000025 or less of the investor's initial wealth, the chance of increasing one's initial fortune tenfold before halving it is 95% or more. However, it takes an average of 294 billion years to achieve this goal, assuming there are 100 draws per year as there are in the Canadian Lotto 6/49.

The conclusion is that except for millionaires and pooled syndicates, it is not possible to use the unpopular numbers in a scientific way to beat the lotto and have high confidence of becoming rich; these aspiring millionaires are also most

likely going to be residing in a cemetery when their distant heir finally reaches the goal.

5. Timing of decisions

After the investment decision is taken, it is necessary to monitor the performance of the portfolio in terms of accumulated capital. This is particularly true in the case where the dynamics of prices have changed since the time of decision. The use of control limits to detect unacceptable or out of control paths is an additional component in the management of the capital accumulation process (MacLean, Zhao and Ziemba, 2005). If the trajectory of capital accumulation is unacceptable, then corrective action is required. When control limits identify significant deviations from expectations, a time for adjustment is identified. The adjustment requires an evaluation of the pricing model and a revision of investment decisions. Using information on prices collected since the last decision, a new price model is developed, an updated investment strategy is calculated (based on the new forecasts and possible financial constraints). Then new control limits are determined to monitor the trajectory of wealth over the upcoming period.

5.1. Control limits

The investment strategy to accumulate wealth is based on estimates of the parameters which drive the asset prices and preferences for accumulated capital. The strategy and the resulting wealth trajectory are very sensitive to estimation errors for the parameters in returns distributions (Kallberg and Ziemba, 1981, 1984, and Chopra and Ziemba, 1993). The estimation errors are far more important than the frequency of the rebalancing of decisions (Rogers, 2001). A natural way to deal with the uncertain direction of a trajectory of the stochastic dynamic wealth process is to set process control limits and to adjust (update estimates for returns and re-solve the growth-security problem) when a limit is reached. This is standard practice in the control of production processes.

In statistical process control, the control limits are based on the mean and standard deviation, with the convention being six-sigma limits. In the capital accumulation problem, the symmetry of deviations from expectations does not hold. Risk aversion makes the investor more sensitive to large losses. The limits can be selected so that they are consistent with risk aversion and the axioms for risk measures. To develop the control limits consider

$\tau_w(X(t)|w_t) =$ the first passage time to wealth w, starting from wealth w_t at time t and following strategy $X(t)$.

Then the upper control limit (UCL) will be set from the expected return at the planning horizon, the level expected if the estimates for model parameters are correct. It is expected that the UCL is reached exactly as the time horizon arrives. The lower control

Table 17
Control limits

$w_U = w_t \exp\{(X'(\phi - re) + r)(T - t)\}$
$w_L = w_t \left(\frac{\alpha_L}{1-(1-\alpha_L)(\frac{w_t}{w_U})^{\theta(X)}}\right)^{\frac{1}{\theta(X)}}$

Source: MacLean, Zhao and Ziemba (2005).

limit (LCL) provides downside risk control. To match the security provided by the VaR problem for example, the LCL is selected so that the wealth process with an optimal strategy will reach the LCL before the UCL at most $100\alpha\%$ of the time.

Suppose the optimal strategy is $X(t)$. The control limits are determined for the wealth trajectory resulting from the decision $X(t)$. Define the UCL as

$$w_U = E\big[W(T)\big]. \tag{57}$$

Then the measure $\delta_2(w_L, w_U)$ can be used to determine the LCL, with

$$w_L = \sup\big\{w \mid \delta_2(w, w_U) \geqslant 1 - \alpha\big\}. \tag{58}$$

The computational formulas for ρ_2 and δ_2 make the calculation of limits straightforward, when the investment decision is a fixed mix such as the fractional Kelly. For the continuous time model the limits have a closed form solution (MacLean, Zhao and Ziemba, 2005) as displayed in Table 17.

To summarize the approach, the growth-security problem is solved at time t for an optimal strategy based on forecast returns and the planning horizon of $(T - t)$. Then control limits are computed, which are consistent with the growth-security specifications, and serve as stopping boundaries for the wealth process. The intention is that the portfolio rebalancing would only take place when a boundary is reached and the trajectory is not proceeding as anticipated.

The control limits can also be used as wealth goals or targets (Bordley and LiCalzi, 2000). Then investment decisions are selected to optimize the time to wealth goals (Burkhardt, 1998). The results are similar to those where the goals are control limits.

5.1.1. Example: Investment in stocks and bonds

The information basis for investment will be daily trading prices for stocks, bonds and cash. These prices will be generated from the random rates of return model. The baseline values for the true price process were presented in Table 8 previously.

The approach to analyzing investment decisions will be to take the VaR model as the standard, see Jorion (1997) and Basak and Shapiro (2001), and to add upper and lower wealth limits calculated from the VaR strategy. That is, at each rebalance time, first the strategy is computed with risk specifications. Then the upper and lower wealth limits will be calculated from the formulas in Table 17. The upper and lower control limits are

Table 18
Control limits (w_L, w_U) for VaR strategies

ρ_1^*	α_L				
	0.01	0.02	0.03	0.04	0.05
0.99	(0.9995, 1.0031)	(0.9996, 1.0033)	(1, 1.0034)	(1, 1.0035)	(1, 1.0037
0.98	(0.9986, 1.0042)	(0.9992, 1.0045)	(0.9999, 1.0047)	(0.9999, 1.0047)	(1,1.0047)
0.97	(0.9976, 1.0047)	(0.9989, 1.0047)	(0.9999, 1.0047)	(0.9999, 1.0047)	(1, 1.0047)
0.96	(0.9976, 1.0047)	(0.9989, 1.0047)	(0.9999, 1.0047)	(0.9999, 1.0047)	(1,1.0047)
0.95	(0.9976, 1.0047)	(0.9989, 1.0047)	(0.9999, 1.0047)	(0.9999, 1.0047)	(1,1.0047)

Source: MacLean, Zhao and Ziemba (2005).

calculated to determine the next rebalance time. The control limits, corresponding to the VaR strategies are shown in Table 18. The initial wealth is $w_0 = 1$, and the VaR horizon is $T = 10$ days. The values in these tables are updated as the portfolio is rebalanced.

The control limits are tight in this example. In all cases the lower limit is above the VaR value ρ_1^*. This is important because the VaR strategy is computed with estimated returns distributions. Although the constraint requires falling below the VaR value at most $100\alpha_L$ percent of the time, the actual trajectories drop below VaR much more frequently—about 4 times as often. The control limits will keep the failures close to the planned $100\alpha_L$ percent. The upper limit hits a ceiling at the optimal growth value of 1.0047.

The values in these tables are updated as the portfolio is rebalanced. The computational experiment which tests the risk control methodology consists of generating daily prices for stocks and bonds for one year (260 trading days), using the lognormal model with parameter values in Table 8. The results for VaR strategies with fixed time rebalancing set at every 10 days were presented in Table 10. In contrast, the results for VaR strategies with random time rebalancing determined by the control limits, where the upper limit is the expected VaR wealth after 10 days, are presented in Table 19.

The average terminal wealth when control limits are included is higher in all the scenarios considered. The advantage grows as the value at risk (fallback) decreases. This improved performance is attributable to rebalancing at the right time, that is when the forecast for returns on assets is clearly in error.

6. Legends of capital growth

The powerful results for the capital growth approach to investing might imply that it is the obvious methodology to use when planning a strategy. However, it is not common in investment planning and the terminology "growth portfolio" does not usually refer to the Kelly strategy. Moreover, it is not even taught in many investment courses. There are, however, notable practitioners of the optimal growth methodology that are centi-

Table 19
Expected wealth with control limits

ρ_1	α_L				
	0.01	0.02	0.03	0.04	0.05
0.99	1.0905	1.0799	1.0730	1.0714	1.0726
0.98	1.1050	1.1024	1.1187	1.1227	1.1270
0.97	1.1395	1.1315	1.1547	1.1708	1.1696
0.96	1.1737	1.1700	1.2059	1.2045	1.2085
0.95	1.2156	1.2227	1.2573	1.2802	1.2853

Source: MacLean, Zhao and Ziemba (2005).

millionaires. Foremost is Edward O. Thorp, who has championed the Kelly strategy in gambling and investment for decades; see his companion chapter that details his experiences and successes and extensive mathematical results on capital growth theory. In this section, some very successful applications of the capital growth strategy are described.

6.1. Princeton Newport Partners

In 1969 an investment partnership specializing in convertible hedging was established, with the plan to use the Kelly strategy to allocate assets. Princeton Newport Partners was managed by Edward O. Thorp, who used the Kelly strategy in the gambling game of Blackjack (Thorp 1962, 1966). To say PNP was successful is an understatement. PNP found risk-arbitrage opportunities in convertible securities that allowed it to get remarkable returns (Thorp and Kassouf, 1967; Thorp, 1969). Figure 8 shows the cumulative earnings results between 1968 and 1988.

Thorp's fund had a net mean return of 15.1% and a yearly standard deviation of 4%. The fund had no losing quarters and only 3 losing months. PNP was closed in 1988. Later Thorp founded two other equally successful funds—Ridgeline Partners (1994–2002) and XYZ (1992–2002).

6.2. Kings College Chest Fund

Another exceptional but more volatile investment record was achieved by the economist John Maynard Keynes. He ran the Kings College Chest Fund from 1927 until his death in 1945. The fund performance is shown in Figure 9.

Keynes lost more than 50% of the fund during the difficult years of the depression. Otherwise the performance was very good but volatile. By 1945, the geometric mean return was 9.12% versus the UK market rate of −0.89%. So Keynes outperformed the market by more than 10% per year. Ziemba (2003) found that The Kings College Chest

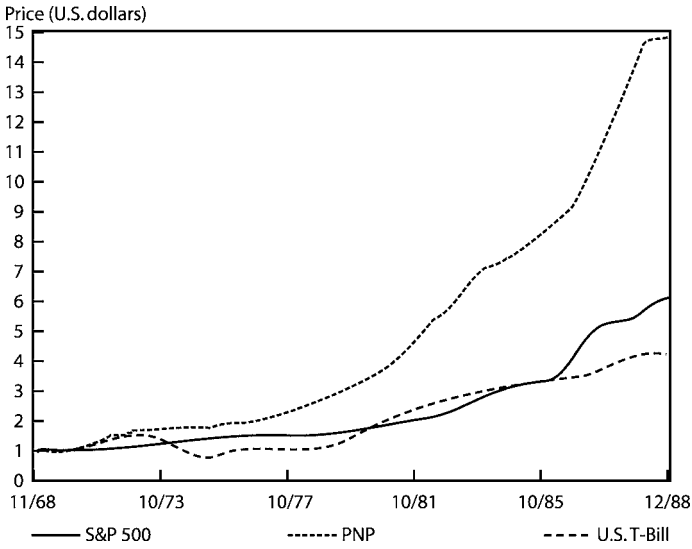


Fig. 8. PNP Performance. Source: Ziemba (2005).

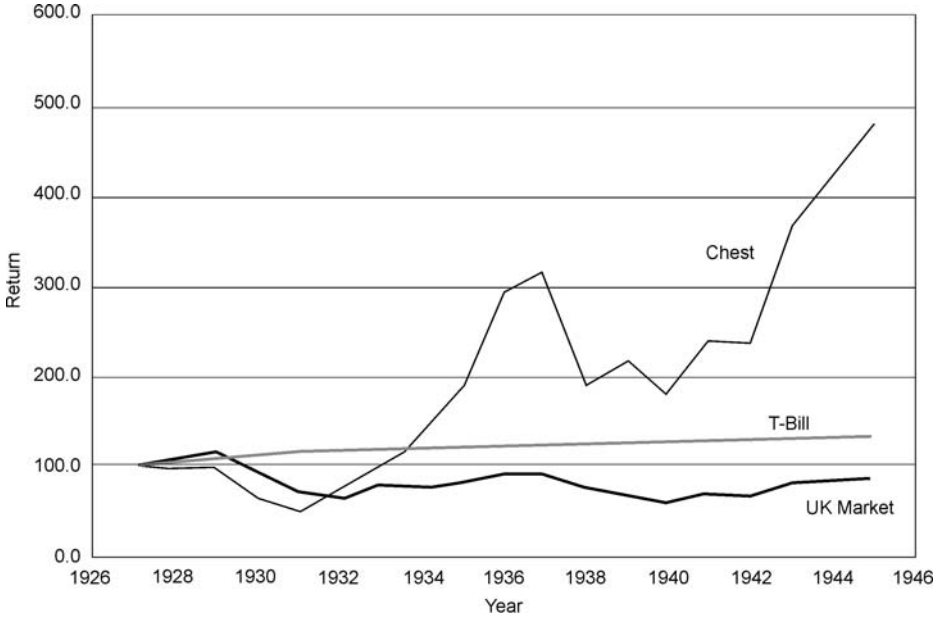


Fig. 9. Chest Fund Performance. Source: Ziemba (2005).

Fund Performance is well approximated by a fractional Kelly investor, with 80% Kelly, 20% cash which is equivalent to the negative power utility function $-w^{-0.25}$.

6.3. Berkshire–Hathaway

The world's most famous investor is Warren Buffett, who runs the Berkshire–Hathaway Fund. Buffett's investment style is aggressive, with emphasis on value, large holdings and patience. The strong performance of that fund is displayed in Figure 10. The geometric mean of BKH for the 40 years from 1965 to 2004 was 22.02% versus 10% for the S&P500, and $15 invested in BKH in 1965 was worth nearly $90,000 in May 2005. The fund outperformed other well known funds, although the wealth trajectory is more variable. Thorp (2006) indicates that Buffett closely follows a Kelly strategy and this could explain the volatility. Using as a risk measure a modified Sharpe ratio which only considers losses, Ziemba (2005) shows that BH is the only fund in the set: {Ford Foundation, Harvard University Endowment, Quantum, Windsor, Tiger, BH}, whose risk measure improves. However, based on the return over the risk free asset per unit of standard deviation, BH was unable to beat the Ford Foundation or the Harvard University Endowment. This confirms that the Kelly capital growth approach must be measured based on long run wealth (i.e., the geometric mean), not on intermediate levels of wealth or their volatility.

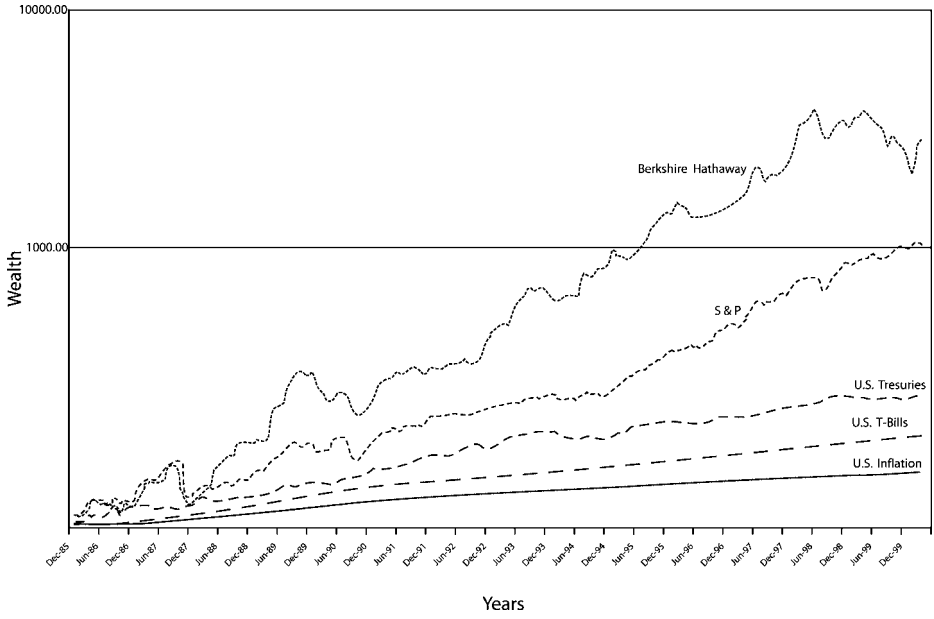


Fig. 10. Berkshire–Hathaway Performance. Source: Ziemba (2005).

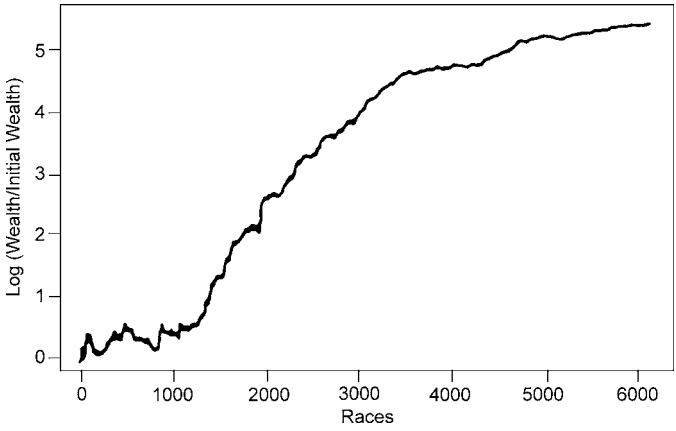


Fig. 11. Hong Kong Betting. Source: Benter (2001).

6.4. *Hong Kong Betting Syndicate*

The world's most successful racetrack bettor is William Benter of the Hong Kong Betting Syndicate (Benter, 1994, 2001). He used a conservative fractional Kelly betting system and other strategies to wager on horse races in Hong Kong over a 12 year period. The performance, as shown in Figure 11, is similar to the pattern of the other growth investors. The growth rate is variable, but averages around 50% per year. One aspect of this application is the importance of accurate estimates of race horse odds and returns. The strong returns after 1000 races correspond to improvements in the handicapping system used by Benter.

One significant feature of the applications described in this section is the long planning horizon. In each case, the early performance was not dominant. But the patience of a long run investor paid off with the capital growth approach. The experience gained over time also led to better estimates of conditional expected returns. As the powerful results of Hens and Schenk-Hoppe (2005) have demonstrated, the capital growth strategy overpowers others when the conditional expected returns are accurate; see also (Breiman, 1961) whose results also suggested this property.

References

Algeot, P., Cover, T., 1988. Asymptotic optimality and asymptotic equipartition Properties of log-optimum investment. Annals of Probability 16, 876–898.

Amir, R., Evstigneev, I., Hens, T., Schenk-Hoppe, K., 2005. Market selection and survival of investment strategies. Journal of Mathematical Economics 41, 105–122.

Artzner, P., Delbaen, F., Eber, J., Heath, D., 1999. Coherent measures of risk. Mathematical Finance 9, 203–228.

Aucamp, D., 1993. On the extensive number of plays to achieve superior performance with the geometric mean strategy. Management Science 39, 1163–1172.
Bain, R., Hausch, D.B., Ziemba, W.T., 2006. An application of expert information to win betting on the Kentucky derby, 1981–2005. European Journal of Finance, in press.
Basak, S., Shapiro, A., 2001. Value at Risk based risk management: optimal policies and asset prices. Review of Financial Studies 14 (2), 371–405.
Bell, R.M., Cover, T.M., 1980. Competitive optimality of logarithmic investment. Mathematics of Operations Research 5, 161–166.
Benter, W., 1994. Computer based horse race handicapping and wagering systems: a report. In: Hausch, D.B., Lo, V., Ziemba, W.T. (Eds.), Efficiency of Racetrack Betting Markets. Academic Press, San Diego, CA, pp. 183–198.
Benter, W., 2001. Development of a mathematical model for successful horse race wagering. Presentation at the University of Nevada, Las Vegas.
Blake, D., Lehmann, B.N., Timmermann, A., 1999. Asset allocation dynamics and pension fund performance. Journal of Business 72, 429–461.
Bordley, R., LiCalzi, M., 2000. Decision analysis using targets instead of utility functions. Decisions in Economics and Finance 23, 53–74.
Breiman, L., 1960. Investment policies for expanding business optimal in a long-run sense. Naval Research Logistics Quarterly 7, 647–651.
Breiman, L., 1961. Optimal gambling system for favorable games. In: Proc. of 4th Berkeley Symposium on Mathematics, Statistics and Probability, vol. 1, pp. 63–68.
Browne, S., 1997. The return on investment from proportional investment strategies. Advances in Applied Probability 30 (1), 216–238.
Browne, S., Whitt, W., 1996. Portfolio choice and the Bayesian Kelly criterion. Advances in Applied Probability 28, 1145–1176.
Burkhardt, T., 1998. A mean-variance of first passage time approach to portfolio selection in a lognormal world. In: Presentation at the VIII International Conference on Stochastic Programming, Vancouver, August.
Chernoff, H., 1980/1981. An analysis of the Massachusetts numbers game. Tech. Rep. No. 23. MIT Department of Mathematics, MIT, Cambridge, MA.
Chernov, M., Gallant, A.R., Ghysels, E., Tauchen, G., 2002. Alternative models for stock price dynamics. Working paper, Columbia Business School, Columbia University.
Chopra, V., Ziemba, W.T., 1993. The effect of errors in the mean, variance, and covariance estimates on optimal portfolio choice. Journal of Portfolio Management, 6–11, Winter.
Clark, R., Ziemba, W.T., 1987. Playing the turn of the year effect with index futures. Operations Research 35, 799–813.
De Giorgio, E., 2005. Reward-risk portfolio selection and stochastic dominance. Journal of Banking and Finance 29 (4), 895–926.
Dohi, T., Tanaka, H., Kaio, N., Osaki, S., 1995. Alternative growth versus security in continuous dynamic trading. European Journal of Operations Research 84, 430–443.
Ethier, S.N., 1987. The proportional bettor's fortune. In: Proceedings 7th International Conference on Gambling and Risk Taking. Department of Economics, University of Nevada, Reno.
Ethier, S.N., Tavare, S., 1983. The proportional bettor's return on investment. Journal of Applied Probability 20, 563–573.
Finkelstein, M., Whitley, R., 1981. Optimal strategies for repeated games. Advanced Applied Probability 13, 415–428.
Gottleib, G., 1985. An analytic derivation of Blackjack win rates. Operations Research 33, 971–988.
Grauer, R.R., Hakansson, N.H., 1986. A half century of returns on levered and unlevered portfolios of stocks, bonds and bills, with and without small stocks. Journal of Business 592, 287–318.
Grauer, R.R., Hakansson, N.H., 1987. Gains from international diversification 1968–85 returns on portfolios of stocks and bonds. Journal of Finance 42, 721–739.

Griffin, P., 1985. Different measures of win rates for optimal proportional betting. Management Science 20, 1540–1547.

Grossman, S.J., Zhou, Z., 1993. Optimal investment strategies for controlling drawdowns. Mathematical Finance 3, 241–276.

Hakansson, N.H., 1970. Optimal investment and consumption strategies under risk for a class of utility functions. Econometrica 38, 587–607.

Hakansson, N.H., 1971. On optimal myopic portfolio policies, with and without serial correlation of yields. Journal of Business 44, 324–334.

Hakansson, N.H., Miller, B.L., 1975. Compound-return mean-variance efficient portfolios never risk ruin. Management Science 22, 391–400.

Hakansson, N.H., Ziemba, W.T., 1995. Capital growth theory. In: Jarrow, R.A., Maksimovic, V., Ziemba, W.T. (Eds.), Finance. North-Holland, Amsterdam, pp. 65–86.

Hanoch, G., Levy, H., 1969. The efficiency analysis of choices involving risk. Review of Economic Studies 36, 335–346.

Hausch, D.B., Lo, V., Ziemba, W.T. (Eds.), 1994. Efficiency of Racetrack Belting Markets. Academic Press, San Diego, CA.

Hausch, D.B., Ziemba, W.T., 1985. Transactions cost, extent of inefficiencies, entries and multiple wagers in a racetrack betting model. Management Science 31 (4), 381–392.

Hausch, D.B., Ziemba, W.T., 1990a. Atbritrage strategies for cross-track betting on major horse races. Journal of Business 63, 61–78.

Hausch, D.B., Ziemba, W.T., 1990b. Locks at the racetrack. Interfaces 20 (3), 41–48.

Hausch, D.B., Ziemba, W.T., 1995. Efficiency of sports and lottery betting markets. In: Jarrow, R., Maksimovic, V., Ziemba, W.T. (Eds.), Finance. In: North-Holland Handbooks in Operations Research and Management Science. North-Holland, Amsterdam, pp. 545–580.

Hausch, D.B., Ziemba, W.T., Rubinstein, M., 1981. Efficiency of the markets for racetrack betting. Management Science 27, 1435–1452.

Hens, T., Schenk-Hoppe, K., 2005. Evolutionary stability of portfolio rules in incomplete markets. Journal of Mathematical Economics 41, 43–66.

Hensel, C.R., Ezra, D.D., Ilkiw, J.H., 1991. The importance of the asset allocation decision. The Financial Analysts Journal, 65–72, July–August.

Hensel, C.R., Ziemba, W.T., 2000. Anticipation of the January small firm effect in the US futures markets. In: Kiem, D., Ziemba, W.T. (Eds.), Security Market Imperfections in Worldwide Equity Markets. Cambridge University Press, Cambridge, UK, pp. 179–202.

Janacek, K., 1998. Optimal growth in gambling and investing. M.Sc. Thesis, Charles University, Prague.

Jorion, P., 1986. Bayes–Stein estimation for portfolio analysis. Journal of Financial and Quantitative Analysis 21, 279–329.

Jorion, P., 1997. Value-at-Risk: The Benchmark for Controlling Market Risk. Irwin, Chicago.

Kallberg, J.G., Ziemba, W.T., 1981. Remarks on optimal portfolio selection. Methods of Operations Research 44, 507–520.

Kallberg, J.G., Ziemba, W.T., 1983. Comparison of alternative utility functions in portfolio selection problems. Management Science 29, 1257–1276.

Kallberg, J.G., Ziemba, W.T., 1984. Mis-specification in portfolio selection problems. In: Bamberg, G., Spreemann, K. (Eds.), Risk and Capital. Springer-Verlag, Berlin/New York, pp. 74–87.

Kelly, J., 1956. A new interpretation of information rate. Bell System Technology Journal 35, 917–926.

Latane, H., 1959. Criteria for choice among risky ventures. Journal of Political Economy 67, 144–145.

Levy, H., 1998. Stochastic Dominance. Kluwer Academic Publishers, Norwell, MA.

Lorenz, M., 1905. Methods for measuring the concentration of wealth. Journal of the American Statistical Association 9, 209–219.

Luenberger, D.G., 1993. A preference foundation for log mean-variance criteria in portfolio choice problems. Journal of Economic Dynamics and Control 17, 887–906.

Luenberger, D.G., 1998. Investment Science. Oxford University Press, New York.

MacLean, L.C., Foster, M.E., Ziemba, W.T., 2002. Empirical Bayes estimation of securities price parameters. In: Paradolos, P. (Ed.), Financial Engineering, E-Commerce and Supply Chain. Kluwer, Boston, MA, pp. 47–57.

MacLean, L.C., Foster, M.E., Ziemba, W.T., 2005. Empirical Bayes estimation with dynamic portfolio models. Technical Report, Sauder School of Business, UBC, Vancouver.

MacLean, L.C., Weldon, K.L., 1996. Estimating multivariate random effects without replication. Communications in Statistics—Theory and Methods 24, 1447–1469.

MacLean, L.C., Zhao, Y., Ziemba, W.T., 2005. Dynamic portfolio selection with process control. Journal of Banking and Finance 30 (2), 317–339.

MacLean, L.C., Zhao, Y., Ziemba, W.T., 2006. Growth-security models and stochastic dominance. In: Dantzig, G.B., Infanger, G. (Eds.), Planning Under Uncertainty. Springer, in press.

MacLean, L.C., Ziemba, W.T., 1991. Growth–security profiles in capital accumulation under risk. Annals of Operations Research 31, 501–510.

MacLean, L.C., Ziemba, W.T., 1999. Growth versus security tradeoffs in dynamic investment analysis. Annals of Operations Research 85, 193–225.

MacLean, L.C., Ziemba, W.T., 2000. Efficiency concepts in capital accumulation models. Philosophical Transactions of the Royal Society A 357, 2029–2037.

MacLean, L.C., Ziemba, W.T., Blazenko, G., 1992. Growth versus security in dynamic investment analysis. Management Science 38, 1562–1585.

MacLean, L., Ziemba, W.T., Li, Y., 2005. Time to wealth goals in capital accumulation. Quantitative Finance 5 (4), 343–357.

MacLean, L.C., Sanegre, R., Zhao, Y., Ziemba, W.T., 2004. Capital growth with security. Journal of Economic Dynamics and Control 28, 937–954.

Markowitz, H.M., 1952. Portfolio selection. Journal of Finance 7, 77–91.

Markowitz, H.M., 1976. Investment for the long run: New evidence for an old rule. Journal of Finance 315, 1273–1286.

Markowitz, H.M., 1987. Mean-Variance Analysis in Portfolio Choice and Capital Markets. Basil Blackwell, New York.

Mossin, J., 1968. Optimal multiperiod portfolio policies. Journal of Business 41, 215–229.

Mulvey, J., Vladimirou, H., 1992. Stochastic network programming for financial planning problems. Management Science 38 (11), 1642–1664.

Ogryczak, W., Ruszczynski, A., 2002. Dual stochastic dominance and related mean-risk models. SIAM Journal of Optimization 13, 60–78.

Pratt, J.W., 1964. Risk aversion in the small and in the large. Econometrica 32, 122–136.

Rendon, J., Ziemba, W.T., 2005. Is the January effect still alive in the futures market? Working paper, Sauder School of Business, UBC, Vancouver, Canada.

Rockafellar, R.T., Uryasev, S., 2000. Optimization of conditional value-at-risk. Journal of Risk 2, 21–41.

Rogers, L.C.G., 2001. The relaxed investor and parameter uncertainty. Finance and Stochastics 5, 131–154.

Rotando, L.M., Thorp, E.O., 1992. The Kelly criterion and the stock market. American Mathematical Monthly 99, 922–931, December.

Rubinstein, M., 1976. The strong case for the generalized logarithmic utility model as the premier model of financial markets. Journal of Finance 31, 551–571.

Samuelson, P.A., 1971. The fallacy of maximizing the geometric mean in long sequences of investing or gambling. Proceedings National Academy of Science 68, 2493–2496.

Stutzer, M., 2003. Portfolio choice with endogenous utility: a large deviations approach. Journal of Econometrics 116, 365–386.

Stutzer, M., 2004. Asset allocation without unobservable parameters. Financial Analysts Journal 60 (5), 38–51.

Thorp, E.O., 1962. Beat the Dealer. Random House, New York, second ed., Vintage, New York, 1966.

Thorp, E.O., 1969. Optimal gambling systems for favorable games. Review of the International Statistical Institute 37, 273–293.

Thorp, E.O., 1971. Portfolio choice and the Kelly criterion. In: Proceedings of the American Statistical Association, Business and Economic Statistics Section, pp. 215–224.

Thorp, E.O., 1975. Portfolio choice and the Kelly criterion. In: Ziemba, W.T., Vickson, R.G. (Eds.), Stochastic Optimization Models in Finance. Academic Press, New York, pp. 599–619. Reprinted from Thorp (1971).

Thorp, E.O., 2006. The Kelly criterion in Blackjack, sports betting and the stock market. In: Zenios, S.A., Ziemba, W.T. (Eds.), Handbook Asset and Liability Management, vol. 1, pp. 385–428, Chapter 9 in this volume.

Thorp, E.O., Kassouf, S.T., 1967. Beat the Market. Random House, New York.

Wilcox, J., 2003. Harry Markowitz and the discretionary wealth hypothesis. Journal of Portfolio Management 29 (3), 58–65.

Wilcox, J., 2005. A better paradigm for finance. Finance Letters 3 (1), 5–11.

Ziemba, W.T., 1994. Investing in the Turn-of-the-Year Effect in the US Futures market. Interfaces 24 (3), 46–61.

Ziemba, W.T., 2003. The Stochastic Programming Approach to Asset, Liability, and Wealth Management. AMIR, Virginia, USA.

Ziemba, W.T., 2005. The symmetric downside-risk Sharpe ratio and the evaluation of great investors and speculators. Journal of Portfolio Management, 108–122. Fall.

Ziemba, W.T., Hausch, D.B., 1984. Beat the Racetrack. Harcourt, Brace and Jovanovich, San Diego, CA.

Ziemba, W.T., Hausch, D.B., 1986. Betting at the Racetrack. Dr. Z Investments, Inc., San Luis Obispo, CA.

Ziemba, W.T., Hausch, D.B., 1987. Dr. Z's Beat the Racetrack. William Morrow, New York. Update of 1984, Beat the Racetrack, Harcourt, Brace and Jovanovich, San Diego, CA.

Ziemba, W.T., Brumelle, S.L., Gautier, A., Schwartz, S.L., 1986. Dr. Z's 6/49 Guidebook. Dr. Z Investments, Inc., San Luis Obispo, CA.

This page intentionally left blank

AUTHOR INDEX

n indicates citation in footnote.

Abaffy, J., *see* Dupačová, J. 58
Adaci, J. 206
Adams, K.J. 36, 43
Aitken, A.C. 143n
Algeot, P. 446
Allen, F. 8
Amir, R. 440, 443, 446
Andersson, F. 13
Andrews, K.R. 3
Androulakis, I.P., *see* Maranas, C.D. 289
Arrow, K.J. 180, 209
Artzner, P. 13, 288, 433
Asay, M.R. 269
Associates First Capital Corporation, 410
Aucamp, D. 446
Aziz, A.R., *see* Dembo, R.S. 13, 19

Babbel, D.F., *see* Santomero, A.M. 15
Bain, R. 460, 461
Baker, N., *see* Haugen, R. 188
Balduzzi, P. 290
Barber, B. 189
Barberis, N. 290
Basak, S. 464
Basel Committee on Banking Supervision, 12, 16
Bawa, V.S. 288
Beale, E.M.L. 256
Beebower, G.L., *see* Brinson, G.P. 187
Bell, D. 210
Bell, R.M. 446
Bellman, R.E. 184, 206
Benter, W. 469
Berger, A.J., *see* Maranas, C.D. 289
Berkelaar, A. 289
Bernoulli, D. 208
Bernstein, P.L. ix
Bertocchi, M. 137, 308, 311, 314, 321, 323
Bertocchi, M., *see* Dupačová, J. 58, 282, 298, 311, 330
Bianchi, S., *see* Wets, R. 38
Bierwag, G. 137
Billingsley, P. 298
Birge, J.R. 21, 256, 308
BIS, 35, 36
Black, F. 9, 58, 143, 188, 270, 271, 310, 316, 407
Blake, D. 201, 455
Blazenko, G., *see* MacLean, L.C. 446, 457–462
Blin, J. 190
Bloch, M. 185
Blume, M., *see* Friend, I. 211
Blume, M.E. 189
Bodie, Z. 211
Bodnar, G.M. 4, 17
Boender, G.C.E. 267, 268, 270, 274, 288, 289
Bordley, R. 464
Borodovsky, L., *see* Lore, M. 10, 15
Bouyoucos, P.J., *see* Asay, M.R. 269
Boyd, S. 38
Brace, A. 58
Bradley, S.P. 255, 292, 311
Breiman, L. 387, 389, 432, 446, 469
Brennan, M.J. 206, 271, 290
Brigo, D. 58
Brinson, G.P. 187
Britt, S., *see* Mulvey, J.M. 206
Brooke, A. 205
Brooks-Hill, R., *see* Ziemba, W.T. 207
Browne, S. 387, 391, 392, 395, 433, 438, 446
Brumelle, S.L., *see* Ziemba, W.T. 461
Buffett, M. 413
Buono, M. 42
Burkhardt, T. 464

Cagan, L.D. 270
Cairns, A.J.G. 290
Campbell, J.Y. 203, 339n, 352n, 365
Capitani, C., *see* Brigo, D. 58
Cariño, D.R. 13, 204, 255, 265–267, 271, 279, 282, 287, 291, 292, 294, 296, 298, 313
Carleton, W., *see* Chambers, D. 137

Carriero, N.S., *see* Cagan, L.D. 270
Censor, Y. 256, 297
Chacko, G. 203, 339n, 367, 371, 372
Chacko, G., *see* Campbell, J.Y. 352n
Chamberlain, G. 181
Chambers, D. 137
Chambers, D., *see* Nawalkha, S. 72, 137
Chan, K. 188
Chernoff, H. 461
Chernov, M. 439
Chopra, V.K. 191, 219, 280, 436, 437, 446, 463
Christensen, P.E. 293
Christofi, C. 171
Clark, D., *see* Buffett, M. 413
Clark, R. 458, 459
Clausing, D., *see* Hauser, J.R. 7
Cocco, F., *see* Consiglio, A. 206, 255, 266
Cohen, K.J. 148, 189
Collomb, A. 205, 247
Connor, G. 222
Consigli, G. 206, 255, 292, 296
Consigli, G., *see* Dupačová, J. 282, 283, 311
Consiglio, A. 8, 206, 255, 265, 266, 273, 292
Constantinides, G.M. 201
Cover, T.M., *see* Algeot, P. 446
Cover, T.M., *see* Bell, R.M. 446
Cox, D.R. 414
Cox, J.C. 203
Crane, D.B., *see* Bodie, Z. 211
Crane, D.B., *see* Bradley, S.P. 255, 292, 311
Crouhy, M. 10

Dai, Q. 341n
Dantzig, G.B. 148, 162, 175, 176, 205, 256, 297
de Boor, C. 86
De Farias, D.P. 206
De Giorgio, E. 433, 452
De Jong, F. 58
de La Grandville, O. 20, 93, 99, 137, 307
Deaves, R. 33
Delbaen, F., *see* Artzner, P. 13, 288, 433
Dembo, R.S. 9, 13, 19, 211, 307
Dempster, M.A.H. 297
Dempster, M.A.H., *see* Consigli, G. 255, 292, 296
Derman, E., *see* Black, F. 58, 270, 271, 310, 316
Dert, C. 255, 266, 278, 292
Dexter, A.S. 183
Dimson, E. 187, 201
Dohi, T. 458
Douglas, G., *see* Blin, J. 190
Driessen, J., *see* De Jong, F. 58
Duarte, A., *see* Koskosides, Y. 271, 276
Duffie, D. 271, 341n, 344
Dupačová, J. 58, 282, 283, 298, 308, 311, 314, 318, 320, 323, 330, 334
Dupačová, J., *see* Bertocchi, M. 311, 314, 321, 323

Eber, J.-M., *see* Artzner, P. 13, 288, 433
Eckstein, J., *see* Hiller, R.S. 204, 292, 307, 311
Ederington, L.H. 183
Edirishinghe, N.C.P. 206
Elton, E.J. 170, 171, 189
Embrechts, P. 13, 268, 270, 277
Epstein, L.G. 344
Epstein, L.G., *see* Duffie, D. 344
Ethier, S.N. 446
Evstigneev, I., *see* Amir, R. 440, 443, 446
Ezra, D.D., *see* Hensel, C.R. 455

Fabozzi, F.J., *see* Christensen, P.E. 293
Faig, M. 211
Fama, E.F. 188, 288
Farrell Jr., J.L. 187
Feeney, G.J. 190
Feller, W. 143n, 180, 387, 388
Finkelstein, M. 446
Fishburn, P.C. 288
Fisher, L. 93, 137
Fisher, M. 344
Fletcher, R. 166
Fleten, S.-E. 292
Floudas, C.A., *see* Maranas, C.D. 289
Fong, H.G., *see* Vasicek, O.A. 43
Foster, M.E., *see* MacLean, L.C. 442–444, 446, 449
Frauendorfer, K. 205, 310, 311
French, K.R., *see* Fama, E.F. 188
Friend, I. 211
Frost, P. 190

Galai, D., *see* Crouhy, M. 10
Gale, D., *see* Allen, F. 8
Gallant, A.R., *see* Chernov, M. 439
Garman, M. 14
Gassmann, H.I., *see* Dempster, M.A.H. 297
Gatarek, D., *see* Brace, A. 58
Gautier, A., *see* Ziemba, W.T. 461
Ghysels, E., *see* Chernov, M. 439
Giacometti, R., *see* Bertocchi, M. 137, 308
Gibson, R.C. 187

Gilles, C., *see* Fisher, M. 344
Giovannini, A. 346
Glynn, P.W., *see* Dantzig, G.B. 205
Goldstein, A.B. 268
Golub, B. 204, 255, 269, 282, 292, 311
Gondzio, J. 267, 286, 294–296
Gordon, M. 187
Gottleib, G. 446
Gould, G., *see* Mulvey, J.M. 255, 271, 292
Grauer, R.R. 184, 432, 443
Green, W.H. 271
Gregoru-Allen, R.B., *see* Buono, M. 42
Griffin, P.A. 399, 404, 446
Griffiths, W.E., *see* Judge, G.G. 276
Grinold, R.C. 222
Grossman, S.J. 445
Gröwe-Kuska, N. 311, 318, 320, 321, 334
Gröwe-Kuska, N., *see* Dupačová, J. 282, 298, 311, 318, 320, 323, 334
Gruber, M.J., *see* Elton, E.J. 170, 171, 189
Guerard, J., *see* Bloch, M. 185
Gunthorpe, D. 201

Hagstrom Jr., R.G. 413
Hakansson, N.H. 20, 21, 184n, 203, 266, 288–290, 340n, 432, 446
Hakansson, N.H., *see* Grauer, R.R. 432, 443
Hamilton, J.D. 271
Hanoch, G. 209, 433, 449
Hansen, L. 271
Harker, P.T. ix, 3
Harlow, W. 288
Harrison, J.M. 284
Hassibi, A., *see* Boyd, S. 38
Haugen, R. 188
Hausch, D.B. 432, 460, 461
Hausch, D.B., *see* Bain, R. 460, 461
Hausch, D.B., *see* Ziemba, W.T. 432, 446, 460
Hauser, J.R. 7
Hawawini, G. 137
Hayt, G.S., *see* Bodnar, G.M. 4, 17
Heath, D. 137
Heath, D., *see* Artzner, P. 13, 288, 433
Hecht, R. 395
Heitsch, H. 282, 298
Heitsch, H., *see* Gröwe-Kuska, N. 311, 318, 320, 321, 334
Hens, T. 432, 445, 446, 469
Hens, T., *see* Amir, R. 440, 443, 446
Hensel, C.R. 455, 459
Hester, D.D., *see* Feeney, G.J. 190
Hicks, J. 137
Hildebrand, F.B. 36
Hill, R.C., *see* Judge, G.G. 276
Hiller, R.S. 204, 292, 307, 311
Hlawitschka, W. 184
Holmer, M.R. 7, 255, 269
Holmer, M.R., *see* Golub, B. 204, 255, 269, 282, 292, 311
Holmer, M.R., *see* Zenios, S.A. 292, 293
Hood, L.R., *see* Brinson, G.P. 187
Hopewell, M. 137
Høyland, K. 255, 267, 271, 280, 281
Høyland, K., *see* Fleten, S.-E. 292
Huang, C., *see* Cox, J.C. 203
Huber, I., *see* Ortobelli, S. 181
Hull, J. 30, 58, 61, 270, 271, 294
Hurt, J., *see* Dupačová, J. 308
Hušková, M., *see* Dupačová, J. 58

Ibbotson Associates, 410, 417
Ilkiw, J.H., *see* Hensel, C.R. 455
Infanger, G. 176, 204, 205, 248
Infanger, G., *see* Collomb, A. 205, 247
Infanger, G., *see* Dantzig, G.B. 205, 297
Ingersoll, J. 137
Institutional Investor, 419

Jacobs, B.I. 165, 168, 170, 188
Jamshidian, F. 58, 272, 308
Janacek, K. 446
Jarrow, R. 137
Jarrow, R., *see* Heath, D. 137
Jobson, J.D. 190
Jobst, N.J. 13, 14, 282
Jorion, P. 12, 177, 442, 464
Judge, G.G. 276

Kahn, R.N. 307
Kahnemann, D. 211
Kaio, N., *see* Dohi, T. 458
Kall, P. 256, 308
Kallberg, J.G. 183, 191, 210, 222, 280, 436, 463
Kamakura Corporation, 17, 20
Kan, R., *see* Duffie, D. 341n
Kang, P. 308
Kang, P., *see* Zenios, S.A. 13, 186
Kassouf, S.T., *see* Thorp, E.O. 391, 418, 466
Kaufman, G., *see* Hopewell, M. 137
Kaut, M., *see* Høyland, K. 281
Keefer, T.E., *see* MacKay, R. 178
Keeney, R. 214

Kelly, J.L. 387, 388, 432
Kendrik, D., *see* Brooke, A. 205
Kent, T., *see* Cariño, D.R. 13, 204, 255, 265–267, 271, 279, 292, 294, 296, 298
Kilpatrick, A. 413
Kim, S., *see* Boyd, S. 38
Kim, T.S. 290, 342, 343
King, A.J., *see* Dembo, R.S. 13
King, B.F. 190
Kingsland, L. 268
Klaassen, P. 267, 284, 286
Klüppelberg, C., *see* Embrechts, P. 13
Kobayashi, K., *see* Konno, H. 288
Konno, H. 13, 175, 288
Koopmans, T.C. 148
Korkie, B., *see* Jobson, J.D. 190
Kortanek, K. 37, 38
Koskosides, Y. 271, 276
Kouwenberg, R. 206, 267, 278, 282, 292, 296, 308
Kouwenberg, R., *see* Berkelaar, A. 289
Kouwenberg, R., *see* Gondzio, J. 267, 286, 294–296
Kreps, D.M., *see* Harrison, J.M. 284
Kritzman, M. 408
Kroll, Y. 183
Kuhn, H.W. 157
Kusy, M.I. 204, 255, 268, 292, 313

Lagnado, R., *see* Brennan, M.J. 206, 271, 290
Lakonishok, J. 188
Lakonishok, J., *see* Chan, K. 188
Lam, J.C. 10, 14–16
Lamont, O. 188
Lanstein, R., *see* Rosenberg, B. 187
Latane, H. 432
Lee, T.C., *see* Judge, G.G. 276
Lehavy, R., *see* Barber, B. 189
Lehmann, B.N., *see* Blake, D. 201, 455
Levy, H. 182, 183, 201, 433
Levy, H., *see* Gunthorpe, D. 201
Levy, H., *see* Hanoch, G. 209, 433, 449
Levy, H., *see* Kroll, Y. 183
Levy, K.N., *see* Jacobs, B.I. 165, 168, 170, 188
Li, Y., *see* MacLean, L.C. 206, 340n, 433, 446
LiCalzi, M., *see* Bordley, R. 464
Lindenberg, E.B., *see* Bawa, V.S. 288
Lintner, J. 142, 143, 169, 188
Litterman, R. 14
Litterman, R., *see* Black, F. 188
Liu, J. 290, 345n
Lo, A.W. 7, 9
Lo, V., *see* Hausch, D.B. 460
Lore, M. 10, 15
Lorenz, M. 451
Louveaux, F., *see* Birge, J.R. 21, 256, 308
Lowenstein, R. 413
Luenberger, D.G. 446
Lütkepohl, U., *see* Judge, G.G. 276
Lynch, A., *see* Balduzzi, P. 290

Macaulay, F. 93, 137
MacKay, R. 178
MacLean, L.C. 206, 340n, 432, 433, 438, 442–446, 449, 452–466
Maranas, C.D. 289
Marciano, A.M., *see* Asay, M.R. 269
Mark, R., *see* Crouhy, M. 10
Markowitz, B.G., *see* Goldstein, A.B. 268
Markowitz, H.M. ix, 10, 17, 20, 142, 143, 143n, 144, 145, 145n, 148, 150, 152, 156, 158, 164, 165, 168, 170–174, 177, 181, 182, 182n, 184–186, 188, 189, 191, 192, 195, 207, 211, 287, 291, 387, 432, 436
Markowitz, H.M., *see* Bloch, M. 185
Markowitz, H.M., *see* Jacobs, B.I. 165, 168, 170
Markowitz, H.M., *see* Kroll, Y. 183
Markowitz, H.M., *see* Levy, H. 182, 183
Marohn, Ch., *see* Frauendorfer, K. 310, 311
Marsh, P., *see* Dimson, E. 187, 201
Marston, R.C., *see* Bodnar, G.M. 4, 17
Mausser, H. 13, 14, 178
Mausser, H., *see* Andersson, F. 13
Mausser, H., *see* Dembo, R. 9, 211
McEnally, R.W. 387
McEnally, R.W., *see* Chambers, D. 137
McKendall, R., *see* Golub, B. 204, 255, 269, 282, 292, 311
McKendall, R., *see* Zenios, S.A. 292, 293
McNichols, M., *see* Barber, B. 189
Medvedev, V., *see* Kortanek, K. 37
Meeraus, A., *see* Brooke, A. 205
Mercurio, F., *see* Brigo, D. 58
Merton, R.C. 9, 143, 202, 290, 339, 346
Messina, E. 313
Miccolis, J. 3
Michaud, R.O. 190, 191, 191n, 222
Mikosch, T., *see* Embrechts, P. 13
Miller, B.L., *see* Hakansson, N.H. 446
Miller, H.D., *see* Cox, D.R. 414
Mitra, G., *see* Messina, E. 313
Morgan, C., *see* Mulvey, J.M. 255, 271, 292

Morgenstern, O., *see* Von Neumann, J. 182, 184, 208
Moriggia, V., *see* Bertocchi, M. 311, 314, 321, 323
Moriggia, V., *see* Dupačová, J. 282, 298, 311, 330
Morin, F., *see* Mulvey, J.M. 206
Morton, A., *see* Heath, D. 137
Morton, D., *see* Infanger, G. 205
Mossin, J. 20, 21, 143, 169, 184n, 203, 266, 288, 289, 446
Mulvey, J.M. 13, 14, 19, 204, 206, 255, 267, 271, 282, 292, 293, 296, 308, 313, 432
Mulvey, J.M., *see* Maranas, C.D. 289
Mulvey, J.M., *see* Ziemba, W.T. x, 14, 21, 176, 255, 287, 292
Musiela, M., *see* Brace, A. 58
Musumeci, J. 206, 211
Musumeci, J., *see* Musumeci, J. 206, 211
Myers, D.H., *see* Cariño, D.R. 13, 204, 255, 265–267, 271, 279, 282, 287, 291, 292, 294, 296, 298

Naik, V. 285
Narasimhan, C., *see* Chan, K. 188
Nawalkha, S. 72, 137
Nelson, C. 35
Neumar, K., *see* Chacko, G. 203
Nielsen, L. 340n
Nielsen, S.S. 204, 255, 292, 296, 308

Ogryczak, W. 433, 451
Ohlson, J.A. 181, 188
Omberg, E., *see* Kim, T.S. 290, 342, 343
Ortobelli, S. 181
Osaki, S., *see* Dohi, T. 458
Owen, J. 180, 181

Padberg, M.W., *see* Elton, E.J. 170, 171
Pakes, A. 118, 128, 137
Parkan, C., *see* Ziemba, W.T. 207
Parker, G., *see* Cairns, A.J.G. 290
Parlar, M., *see* Deaves, R. 33
Pauling, B., *see* Mulvey, J.M. 206
Pelsser, A., *see* De Jong, F. 58
Penman, S.H. 188
Perold, A.F. 19, 156
Perold, A.F., *see* Markowitz, H.M. 148, 164, 165
Pflug, G.Ch. 267, 283, 298, 321
Pogue, J.A., *see* Cohen, K.J. 148, 189
Pohlman, L., *see* Golub, B. 204, 255, 269, 282, 292, 311
Porter, M.E. 3
Pratt, J.W. 180, 209, 432
Prékopa, A. 308
Pulley, L.M. 183

Quaife, A. 407

Rabinovitch, R., *see* Owen, J. 180, 181
Raiffa, H., *see* Keeney, R. 214
Reddington, F.M. 20, 137, 293
Reid, D.W., *see* Markowitz, H.M. 182n
Reid, K., *see* Rosenberg, B. 187
Rendon, J. 460
Riskmetrics, 272
Rockafellar, R.T. 13, 180, 288, 451
Rodriguez, J., *see* Campbell, J.Y. 352n
Rogers, L.C.G. 433, 438, 446, 463
Römisch, W., *see* Dupačová, J. 282, 298, 311, 318, 320, 323, 334
Römisch, W., *see* Gröwe-Kuska, N. 311, 318, 320, 321, 334
Römisch, W., *see* Heitsch, H. 282, 298
Ronn, E.I., *see* Nielsen, S.S. 308
Rosen, D., *see* Andersson, F. 13
Rosen, D., *see* Dembo, R.S. 13, 19
Rosen, D., *see* Mausser, H. 13, 14, 178
Rosenberg, B. 187, 190
Ross, S. 190
Rotando, L.M. 391, 432
Roy, A.D. 142, 177
Rubinstein, M. 210, 432
Rubinstein, M., *see* Hausch, D.B. 432, 460
Rudolf, M. 290
Ruszczynski, A., *see* Ogryczak, W. 433, 451

Samuelson, P.A. 137, 181, 184n, 202, 266, 288–290, 446
Sanegre, R., *see* MacLean, L.C. 432, 445, 446, 456, 457
Santomero, A.M. 15
Satchell, S., *see* Sortino, F. 175
Saunders, D., *see* Dembo, R. 9
Savage, L.J. 184, 188
Savarino, J., *see* Frost, P. 190
Schaack, C., *see* Hiller, R.S. 307
Scheinkman, J., *see* Hansen, L. 271
Schenk-Hoppe, K., *see* Amir, R. 440, 443, 446
Schenk-Hoppe, K., *see* Hens, T. 432, 445, 446, 469
Scholes, M.J., *see* Black, F. 9, 407

Schwartz, E.S., *see* Brennan, M.J. 206, 271, 290
Schwartz, E.S., *see* Ortobelli, S. 181
Schwartz, S.L. 188
Schwartz, S.L., *see* Ziemba, W.T. 461
Shah, S., *see* Miccolis, J. 3
Shapiro, A., *see* Basak, S. 464
Shapiro, J.F. 307, 311
Sharpe, W.F. 13, 17, 20, 142, 143, 147, 168–170, 188, 189, 287
Sharpe, W.F., *see* Perold, A.F. 19
Shea, G.S. 36, 43
Shleifer, A., *see* Lakonishok, J. 188
Shtilman, M.S. 267, 282
Shtilman, M.S., *see* Zenios, S.A. 317, 322
Siegel, A., *see* Nelson, C. 35
Siegel, J. 201
Simaan, Y. 183
Singer, B.D., *see* Brinson, G.P. 187
Singleton, K., *see* Duffie, D. 271
Singleton, K.J., *see* Dai, Q. 341n
Skelton, M., *see* Ingersoll, J. 137
Sortino, F.A. 175, 288
Stacy, C., *see* Cariño, D.R. 13, 204, 255, 265–267, 271, 279, 292, 294, 296, 298
Starer, D., *see* Jacobs, B.I. 170
Staunton, M., *see* Dimson, E. 187, 201
Štěpán, J., *see* Dupačová, J. 308
Stulz, R. 10, 14
Stutzer, M. 446–448, 451
Svensson, L. 35
Swensen, D.F. 187
Sylvanus, M., *see* Cariño, D.R. 13, 204, 255, 265–267, 271, 279, 292, 294, 296, 298

Tanaka, H., *see* Dohi, T. 458
Tauchen, G., *see* Chernov, M. 439
Tavare, S., *see* Ethier, S.N. 446
Tew, B.V., *see* Markowitz, H.M. 182n
Theodossiou, P., *see* Christofi, C. 171
Thorlacius, E., *see* Mulvey, J.M. 308
Thorp, E.O. 387, 389, 391, 399, 418, 419, 432, 446, 466, 468
Thorp, E.O., *see* Rotando, L.M. 391, 432
Timmermann, A., *see* Blake, D. 201, 455
Tint, L.G., *see* Sharpe, W.F. 17, 20, 147, 287
Tobin, J. 142, 143, 166, 180, 207
Todd, P., *see* Bloch, M. 185
Todd, P., *see* Markowitz, H.M. 145n, 148, 150, 156, 158, 165, 174, 177
Topaloglou, N. 265
Toy, W., *see* Black, F. 58, 270, 271, 310, 316
Trent, R.H., *see* Young, W.E. 181, 182
Treynor, J.L. 202
Trueman, B., *see* Barber, B. 189
Tucker, A.W., *see* Kuhn, H.W. 157
Tucker, H. 422
Turnbull, S., *see* Jarrow, R. 137
Turner, A.L., *see* Cariño, D.R. 13, 204, 255, 265–267, 271, 279, 292, 294, 296, 298, 313
Tversky, A., *see* Kahnemann, D. 211

Uryasev, S., *see* Andersson, F. 13
Uryasev, S., *see* Rockafellar, R.T. 13, 180, 288, 451
Usmen, N., *see* Markowitz, H.M. 192

van der Meer, R., *see* Sortino, F.A. 288
Van Deventer, D.R., *see* Adams, K.J. 36, 43
van Dijk, E., *see* Markowitz, H.M. 184, 207
Van Roy, B., *see* De Farias, D.P. 206
Vandenberghe, L., *see* Boyd, S. 38
Vasicek, O.A. 43
Vassalou, M., *see* Nielsen, L. 340n
Vassiadou-Zeniou, C. 308
Vassiadou-Zeniou, C., *see* Worzel, K.J. 186, 288, 292
Vassiadou-Zeniou, C., *see* Zenios, S.A. 292, 293
Viceira, L.M. 339n
Viceira, L.M., *see* Campbell, J.Y. 203, 339n, 352n, 365
Viceira, L.M., *see* Chacko, G. 339n, 367, 371, 372
Vickson, R.G., *see* Ziemba, W.T. x, 21, 255
Vishny, R.W., *see* Lakonishok, J. 188
Vladimirou, H., *see* Mulvey, J.M. 204, 255, 292, 296, 432
Vladimirou, H., *see* Topaloglou, N. 265
Von Neumann, J. 182, 184, 208
Vorst, A.C.F., *see* Gondzio, J. 267, 286, 294, 295

Wachter, J. 345n
Wahba, G. 86
Walden, W., *see* Thorp, E.O. 391
Wallace, S.W. x, 205, 308
Wallace, S.W., *see* Dupačová, J. 282, 283, 311
Wallace, S.W., *see* Fleten, S.-E. 292
Wallace, S.W., *see* Høyland, K. 267, 271, 280, 281
Wallace, S.W., *see* Kall, P. 256, 308
Watanabe, K., *see* Cariño, D.R. 13, 204, 255, 265–267, 271, 279, 292, 294, 296, 298
Weil, Ph., *see* Giovannini, A. 346

Weil, R., *see* Fisher, L. 93, 137
Weil, R., *see* Ingersoll, J. 137
Weldon, K.L., *see* MacLean, L.C. 438
Wets, R.J.-B. 38, 256, 298
White, A., *see* Hull, J. 58, 270, 271
Whitley, R., *see* Finkelstein, M. 446
Whitt, W., *see* Browne, S. 433, 438
Wilcox, J. 447, 455
Williams, J.B. 187
Wind, Y. 8
Winklevoss, H.E. 268
Wong, S. 399
Worzel, K.J. 186, 288, 292

Xia, Y. 290
Xu, G., *see* Bloch, M. 185
Xu, G., *see* Markowitz, H.M. 148, 165, 174

Yaari, U., *see* Buono, M. 42
Yamane, Y., *see* Markowitz, H.M. 148, 165, 174
Yamazaki, H., *see* Konno, H. 13, 175, 288
Yang, L., *see* Wets, R. 38
Young, M.R. 186
Young, W.E. 181, 182
Yu, J.N.W., *see* Dexter, A.S. 183
Yunianto, H., *see* Kortanek, K. 37, 38

Zenios, S.A. ix, 13, 14, 20, 21, 137, 186, 204, 255, 267, 269, 270, 282, 287, 292, 293, 307, 317, 322
Zenios, S.A., *see* Bertocchi, M. 137, 308
Zenios, S.A., *see* Cagan, L.D. 270
Zenios, S.A., *see* Censor, Y. 256, 297
Zenios, S.A., *see* Consiglio, A. 8, 206, 255, 265, 266, 273, 292
Zenios, S.A., *see* Golub, B. 204, 255, 269, 282, 292, 311
Zenios, S.A., *see* Harker, P.T. ix, 3
Zenios, S.A., *see* Holmer, M.R. 7, 255
Zenios, S.A., *see* Jobst, N.J. 13, 14, 282
Zenios, S.A., *see* Kang, P. 308
Zenios, S.A., *see* Kouwenberg, R. 206, 308
Zenios, S.A., *see* Mulvey, J.M. 13, 14, 19, 267, 282, 293, 313
Zenios, S.A., *see* Nielsen, S.S. 204, 255, 292, 296
Zenios, S.A., *see* Shtilman, M.S. 267, 282
Zenios, S.A., *see* Topaloglou, N. 265
Zenios, S.A., *see* Vassiadou-Zeniou, C. 308
Zenios, S.A., *see* Worzel, K.J. 186, 288, 292
Zerbs, M., *see* Dembo, R.S. 13, 19
Zhao, Y., *see* MacLean, L.C. 206, 432, 433, 445, 446, 452–457, 463–466
Zhou, Z., *see* Grossman, S.J. 445
Zhu, Y., *see* Jamshidian, F. 272, 308
Ziemba, W.T. ix, 14, 21, 176, 181, 204, 206, 207, 209, 255, 287, 292, 432, 446, 459–461, 466–468
Ziemba, W.T., *see* Bain, R. 460, 461
Ziemba, W.T., *see* Cariño, D.R. 13, 204, 255, 265–267, 271, 279, 282, 287, 291, 292, 294, 296, 298
Ziemba, W.T., *see* Chopra, V.K. 191, 219, 280, 436, 437, 446, 463
Ziemba, W.T., *see* Clark, R. 458, 459
Ziemba, W.T., *see* Dexter, A.S. 183
Ziemba, W.T., *see* Hakansson, N.H. 340n, 432
Ziemba, W.T., *see* Hausch, D.B. 432, 460, 461
Ziemba, W.T., *see* Hensel, C.R. 459
Ziemba, W.T., *see* Kallberg, J.G. 183, 191, 210, 222, 280, 436, 463
Ziemba, W.T., *see* Kusy, M.I. 204, 255, 268, 292, 313
Ziemba, W.T., *see* MacLean, L.C. 206, 340n, 432, 433, 442–446, 449, 452–466
Ziemba, W.T., *see* Rendon, J. 460
Ziemba, W.T., *see* Rudolf, M. 290
Ziemba, W.T., *see* Schwartz, S.L. 188
Ziemba, W.T., *see* Wallace, S.W. ix, 205, 308
Zin, S., *see* Epstein, L. 344

This page intentionally left blank

SUBJECT INDEX

absolute risk aversion, 180, 183
adaptive models, 257
all feasible $E\sigma$ combinations, 152
anomalies, 187
anomalous effects, naive, 188
anomalous effects, pure, 188
anticipative models, 256
approximation theory, 27, 39
APT factor model, 190
arbitrage, 44, 283
Arbitrage Pricing Theory, 190
arbitrage-free event trees, 283–287
arborescent, 309
ask prices, 263
asset allocation, 292
asset and liability management, 17, 19
asset classes, 187
asset liability modeling, 146, 147
asset prices, 435, 436, 439, 440
asset returns, 270, 271, 274
asynchronous simulation analysis, 168

Bayes estimator, 440, 442
Bayes–Stein estimator, 442, 443
Bayesian, 435
Bayesian dynamic model, 433
Bayesian inference, 192
benchmark portfolio, 149
Berkshire–Hathaway, 468
betting, 386, 401–403, 405, 412
bid prices, 263
binomial interest rate tree, 61
Black–Litterman estimation process, 188
bonds, 73, 219
bootstrapping, 30, 219, 272
borrowing, 146, 312
Brownian motion, 438, 439
– geometric, 453
Business risks, 5

capital accumulation, 433–436
capital asset pricing model, 143, 168, 447
capital growth, 386, 407, 408, 432
CARA, 210
cash, 219
certainty equivalent, 436
certainty equivalent wealth, 208, 237
Chicago Mercantile Exchange, 46
choice of criteria, 185, 186
closed-form solution, 339, 340, 371, 372, 382, 383
conditional value at risk, 143, 180
constant correlation model, 189
constraint, nonanticipativity, 309
constraint matrix, 145
constraints, 312, 313
constraints equality inequality, 144
constraints in optimization, 142, 145, 148, 191
consumption/investment problem, dynamic, 339–341, 345
consumption/investment problem, finite horizon, 349–351, 376–378
consumption/investment problem, infinite horizon, 352, 353, 367–376
contamination, 325
contamination, bounds, 330
contamination, parameter, 330
continuously compounded horizon rate of return, 91
continuously compounded spot rate, 72, 75, 76
continuously compounded total return, 72, 77, 79
control limits, 463, 464
control limits on accumulated capital, 436
corner portfolios, 149, 150
corporate governance, 10, 15
correlation between asset returns, 438
covariance, estimation of, 189
covariance, models of, *see* factor model, scenario model, historical covariances, 165
covariance matrix, 280
credit risks, 5
critical line algorithm, 150, 156–158
critical line algorithm stopping before $\lambda_\varepsilon = 0$, 166
critical lines, 149, 150
CRRA, 209
CVaR, *see* conditional value at risk, 180

DECIS, 205, 248
decision, first-stage, 312

decision, second-stage, 312
decision rule, 289
decreasing relative risk aversion, 211, 213, 228
deposits, 30
derivative, directional, 330
derived utility function, 184
design of financial products, 7
designing the product, 8
directional derivative, 330
discrete-time models, 288
discretionary wealth, 455
discretionary wealth hypothesis, 447
discretization, quarterly, 323
disentanglies anomalies, 188
distance, Kantorovich, 320
dividend discount model, 187
downside risk, 211, 241, 287
dummy slack variables, 159
duration, 71, 313
duration, matching, 307
dynamic games, 293
dynamic investment games, 184
dynamic optimization problem, 216
dynamic portfolio choice, 214
dynamic programming recursion, 216

e, 80
economic factors, 270, 274
effective solutions, 457
efficient EV and $E\sigma$ combinations, 150–152
efficient frontier mean-probability of loss, 177
efficient frontier mean-semideviation, 174, 175
efficient frontier mean-value at risk, 177
efficient frontier mean-variance, 142, 144, 149–152
efficient segment, 157
EGP, *see* Elton, Gruber, Padberg, 170
EGP-like algorithm, 171–173
elliptical distributions, 180
Elton, Gruber, Padberg algorithm, 170
enterprise risk management, 3, 4, 10, 15
– conceptual framework, 6
EpiCurves, 38, 40–42, 46, 48, 51
EpiManager, 65
EpiVolatility, 57, 58, 60, 61
equivalent mean-variance model, 145
estimates of covariance, 189
estimation error, 219
estimation of expected returns, 187–189
Euler–Poisson equation, 86
Eurodollar deposits, 31, 46
Eurodollar futures, 31, 46
event trees, 267, 277
evolutionary stable strategy, 445
expectation, hypothesis, 315
expected loss, 185
expected returns, time-varying, 349–353
expected utility, 208
exponential utility function, 209, 222

factor models, 147, 148, 162–164
fast method, 217
Federal Reserve, 46
financial institutions, 3
financial product design, 7
Financial risk, 4
financial valuation, 270
fixed income, 269, 292
fixed mix, 445
fixed-mix strategies, 234
flat volatilities, 60
forward rate, 89
forward–forward volatility, 61
forward-rate agreements, 30
fractional Kelly strategies, 432, 445, 447, 449, 451, 453–455, 459, 468, 469
function theory, 37
fund-separation, 290
funding the product, 9
futures, 30

gambler's ruin, 461
GAMS, 205
Gaussian, 438
general mean-variance model, 144–146, 156–158
general utility, 211
geometric Brownian motion, 453
geometric mean, 181
geometric programming, 37
good and bad properties of the optimal growth strategy, 446
GP modeling, 38
gradient, 117
growth-security trade-off, 457, 460

HARA, 203, 210
hedging demand, 345, 373
hedging downside risk, 238
Hessian matrix, 112
high-performance computations, 270
historical covariance, 165, 189
historical models, 148

Hong Kong Betting Syndicate, 469
horizon, 310

immunization, 71, 73, 90, 307
immunization parameter, 126
immunization theorem, 102, 109
immunizing portfolio, 116
IN and OUT variables, 155, 156
in-sample returns, 215
increasing absolute risk aversion, 183
increasing relative risk aversion, 213, 226
industry factors, 190
initial portfolio, 224, 228, 234, 241
instantaneous forward rate, 72, 73, 75
insurance, 269
insurance products, 292
interest rate, Black–Derman–Toy, 316
interest rate, short term, 307
interest rate cap, 58
interest rate floor, 58, 60
interest rates, time-varying, 378–382
intertemporal substitution, general elasticity, 353–366
intertemporal substitution, unit elasticity, 345–353
investment example, 222
investment horizon, 201, 214
investment opportunity set, time-varying, 339, 341–344
isomean line, 155

January turn-of-the-year effect, 458, 460

Kelly criterion, 386, 387, 391, 392, 398, 399, 405, 419, 420, 432, 445, 446
Kelly strategy, 447, 448
Kentucky Derby, 460, 461
Kings College Chest Fund, 466, 467
kurtosis, 281

large cap stocks, 458
lattice, binomial, 316
lending, 312
liabilities
– scenarios for, 267
liability hedging, 287
LIBOR, 42
LIBOR market model, 58
life-time portfolio and consumption, 202
linear programming, 45, 159–161
Liquidity risks, 5
logarithmic utility, 386, 432, 436, 450
logarithmic utility function, 203, 209
lognormal asset returns, 432
long/short models, 165
Lorenz curve, 451
lotto games, 461, 463
lower partial moment, 211

Macaulay, 93
Macaulay duration, 93
MAD, *see* mean absolute deviation, 175
management of a business portfolio, 9
many-factor model, 190
Market risks, 5
maximum smoothness, 36, 48
mean, 280
mean absolute deviation, 143, 175, 185
mean reversion, 461
mean-variance analysis, 207
mean-variance approximations to expected utility, 181
minimum guarantee, 292
models, 19
modern portfolio theory, 207
moment of order k of a bond portfolio, 106
Monte Carlo, 61
Monte Carlo pre-sampling, 205
Monte Carlo sampling, 205, 215
monthly rebalancing, 239
mortgage-backed securities, 293
multi-period investment problem, 214
multi-stage stochastic programming, 204, 247
multifactor models, 148
multiperiod, 21, 310
multiperiod, two-stage, 313
multistage, 310
multistage models, 260, 314
multivariate normal variables, 273
myopic, 432
myopic demand, 345, 371
myopic policies, 289
myopic portfolio strategies, 203, 446

naive anomalous effects, 188
negative power utility function, 432, 450, 468
Nelson–Siegel model, 34, 51
net present value, 43
network optimization, 296
nodal, 309
noisy forecast, 222
normally distributed returns, 209

normative analysis, 143
numerical results, 371, 372

one-factor model, 147, 170, 189
one-switch utility functions, 210
operational risks, 5
optimization constraints, 142, 145, 148, 191
optimization time, 246
option hedging, 294
out-of-sample returns, 215, 226, 228, 232, 234, 293

parameter estimation error, 438, 463
parameter uncertainty, 190
path-dependent, 270
pension funds, 268, 274
performance measurement, 10, 14, 15
perturbation method 1, around unit elasticity, 353–360, 369, 370, 372–374, 380, 381
perturbation method 2, around mean consumption/wealth ratio, 360–366, 370, 371, 376, 381, 382
piecewise absolute risk aversion, 212
piecewise exponential utility function, 214
piecewise linear, 44
portfolio, bond, 306
portfolio, buy-and-hold, 315
portfolio, rebalancing, 311
portfolio allocation problems, dynamic, 338, 382
portfolio allocations, 417
portfolio management
– with stochastic programming, 261
portfolio mean, 144
portfolio of products, 11
portfolio variance, 144
positive analysis, 143
post-modern portfolio theory, 207
posynomial, 38
power utility function, 209
precision, relative, 321
pricing, 7
pricing the product, 9
Princeton Newport Partners, 466, 467
principle of optimality, 339, 341, 342, 354, 363
probability distribution, discrete, 308
probability of loss, 143, 185
probability of loss, Gaussian case, 176, 177
probability of loss, non-Gaussian case, 178, 179
probability spaces, 298
prospect theory, 211
pure anomalous effects, 188

quadratic form, 117
quadratic penalty of under-performing, 232
quadratic programming, 166
quadratic utility functions, 211
quarterly rebalancing, 239

random sampling, 277
– adjusted, 279
rank deficient constraint matrix, 160
rational action under risk and uncertainty, 184, 192
recourse models, 258
recursive preferences, 340, 382
replicating strategies, 283
resampled efficient frontier, 191
restricted autoregressive process, 217
risk, neutral, 315
risk aversion, 339, 344, 345, 371, 432, 436
risk aversion coefficient, 209
risk aversion index, 449
risk control, 433
risk free asset, 434
risk management, 10, 13, 14
risk measurement, 10, 12, 13
risk measures, 173–180
risk tolerance, 209, 437
risk-free asset, 453
risk-free return, 207
robustness, test, 324
rolling forward, 313
rolling horizon, 293
Roy's constraint set, 142, 149
Roy's model, 142, 169, 177

safety first, 142
sample, pooled, 330
sampling, nonrandom, 317
scenario, 259, 267
scenario, backward algorithm, 321
scenario, construction, 324
scenario, fast backward algorithm, 323
scenario, forward algorithm, 321
scenario, input structure, 334
scenario, interest rate, 310
scenario, out-of-sample, 333
scenario, reduction, 318
scenario, reduction algorithm, 321
scenario, SCENRED, 322
scenario, tree construction, 321
scenario, what-if analysis, 324
scenario models, 148, 164, 165
scenarios, 308, 456, 465
security analyst recommendations, 189

semideviation, 143, 173
semivariance, 173, 185
separation theorem, not true in general, 169
separation theorem, Sharpe, 168, 169
separation theorem, Tobin, 166–168
serial dependency of asset returns, 216
Sharpe ratio, 168, 342, 347–349, 355–357, 361, 362
Sharpe–Lintner CAPM, 168
short sales, 165
simplex algorithm, 159–161
simulation analysis, asynchronous, 168
single-factor model, 147, 170, 189
single-period, 20
single-period investment problem, 209
single-period portfolio choice, 207
skewness, 281
slack variables, 145
small cap stocks, 458
solution algorithms, 296
solution methods, 296
spherically symmetric distributions, 181
spot volatilities, 60
stability, quantitative, 320
stages, 310
standard constraint set, 142, 150
starting the critical line algorithm, 158
state-dependent, 270
static models, 20, 21
stochastic, programming, 308
stochastic control problem, 339, 382
stochastic dominance, 433, 435, 449–451, 454
stochastic dynamic processes, 434
stochastic dynamic programming, 206
stochastic models, 20, 21
stochastic programming, 176, 256–261
– basics, 256
– deterministic equivalent, 259
– for asset and liability management, 291
stochastic programs, 255
stochastic volatility, 294
stocks, 219
strategic planning, 35
stress, test, 324
subtrees, 324
Svensson model, 34, 51
swap market model, 58
swaps, 30, 31, 46

tangent portfolio, 142, 167
target wealth, 211
term structure, 27, 30, 48, 52
three-factor model, 188
three-security examples, 149–156
time diversification, 201
Tobin's model, 143
topdown analysis, 187
tracking error, 148, 149, 186
trading strategy, 434
transaction costs, 294
treasury bills, 30
treasury bonds, 30
treasury notes, 30
tree construction, 267
tree fitting, 280
trimable model, 165
truncation estimator, 442, 443
turnover constraint, 148
two fund separation theorem, 169, 181, 207

uncertainty, 255
unpopular numbers, 461
upper bounds, 161
utility, expected, 312
utility functions, 209, 339, 340, 344, 345, 349
utility functions, pathologically risk averse, 182

validation, 310
value at risk (VaR), 143, 272
value at risk, Gaussian case, 176, 177
value at risk, non-Gaussian case, 178, 179
value of a blank check, 182
VaR strategies, 455, 457, 465
vector autoregressive process, 216
volatility, 436, 439
volatility, time-varying, 367–378
volatility curve, 61
volatility structure, 57

Warren Buffett, 468
wealth distributions, 451
wealth goals, 433
wealth limits, 464
WealthiOR, 246
world wide asset allocation, 432

yearly rebalancing, 239
yoyo-strategies, 290

zero-curve, 27

This page intentionally left blank