John B. Guerard, Jr.
Editor

Handbook of Portfolio Construction

Contemporary Applications of Markowitz Techniques

Springer

Handbook of Portfolio Construction

John B. Guerard, Jr.
Editor

Handbook of Portfolio Construction

Contemporary Applications of Markowitz Techniques

Springer

John B. Guerard, Jr.
McKinley Capital
Management, LLC
3301 C. Street
Anchorage, AK 99503
USA
jguerard@mckinleycapital.com

ISBN 978-0-387-77438-1 e-ISBN 978-0-387-77439-8
DOI 10.1007/978-0-387-77439-8
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2009933258

© Springer Science+Business Media, LLC 2010
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

I am deeply honored by the articles that appear in this collection. In particular, I thank John Guerard for organizing this recognition of my work. I have thought about whether I should individually thank particular authors. I hesitate to do so because a complete list would be too long, and an incomplete list would leave out names that should not be omitted. I do not suppose, however, that anyone's feelings will be hurt if I specifically thank Paul Samuelson for his contribution. I am sure that the other contributors to this volume – including old friends and colleagues as well as other notable contributors to finance practice and literature – have the same admiration for Paul for what he has to say and how he says it. To Paul in particular, and to each of the other contributors in general, thank you.

The remainder of this note is devoted to points about my work that I would very much like the reader to "please note." For example, many writers seem to know only this about my work: in 1952, I wrote an article that defined and recommended mean-variance analysis. They seem to be unaware that there is a large difference between the views I held in 1952 and those I held and expressed in 1959. Of course, the definition of mean-variance efficiency and the formulas for portfolio mean and portfolio variance are the same in the two works, but the justification for mean-variance analysis, the relationship between single-period and many-period analysis, and the method of computing efficient sets for large numbers of securities are all new in both my1956 article and my 1959 book.

Many ascribe assumptions underlying mean-variance analysis to me; they are, in fact, credited to Tobin (1958) and eschewed by Markowitz (1959). Many contrast the rational decision making of the mean-variance optimizer with the economic actor described in behavioral finance, and assume that I am uninterested in or in opposition to the latter. In fact, while my portfolio selection paper, referenced here as Markowitz (1952a), is generally noted as establishing me as "the father of portfolio theory," my Markowitz (1952b) article is recognized by leading behavioral finance specialists to share a "grandfatherly relationship" with behavioral finance.

Below, I flush out the distinctions between work that I did at different times, on one subject or another, between the views I have expressed and those that are often attributed to me. Each section below has a title such as "Markowitz (1952a) and Markowitz (1952b)." The section title should be thought of as following the phrase "Please note the difference between. . . "

Markowitz (1952a) and Markowitz (1952b)

The "Utility analysis of choices involving risk" by Friedman and Savage (1948) was an assignment in which I took Friedman's course around 1949. I found some difficulties with the assigned paper. My 1952b paper, "The utility of wealth," details these difficulties. At first, the latter paper attracted some but not much attention. Then, Kahneman and Tversky (1979) published their paper on prospect theory. This is referred to as Markowitz (1952b) in a footnote, because the idea of measuring utility as a deviation from current wealth has been used by me as well as by Kahneman and Tversky.

The three-volume handbook by Shefrin (2001) reprints a copy of "The utility of wealth" as the first article in its historical review volume. In this volume, Shefrin describes the nature and role of "The utility of wealth" as follows:

> Markowitz (1952) truly qualifies as a behavioral work, with its focus on how people actually behave. Markowitz addresses a classic question posed by Friedman and Savage (1948): why do people simultaneously purchase insurance and lottery tickets? Friedman and Savage proposed a solution to this question, a solution that Markowitz criticized on behavioral grounds. In arguing against the Friedman–Savage solution, Markowitz described the results of how people behave, citing an experiment conducted by Mosteller and Nogee (1951) about how people bet. Reliance on experimental data fell out of fashion for a while, and still engenders some controversy among financial economists.
>
> Looked at in hindsight, Markowitz showed amazing insight. The theory he proposes as an alternative to Friedman and Savage contains basic elements that were later developed much more fully. His discussion about the difference between present wealth and customary wealth gave rise to the coding of gains and losses relative to a reference point. He recognized that losses loom larger than gains. He proposed a utility function with three inflection points to capture the idea that attitude or risk varied with the situation being faced. In this respect he emphasized the importance of whether a gamble is framed in terms of gains or losses, as well as whether the stakes are small or large. His discussion touches on aspiration points, the preference for positive skewness, and a property Thaler and Johnson (1990) subsequently called the "house money effect."
>
> The ideas introduced in Markowitz (1952) were later developed in prospect theory, a framework proposed by psychologists Kahneman and Tversky (1979). Prospect theory combines the insights of Markowitz with those of Allais (1952). It draws on Markowitz for the concepts of framing, gains, losses, reference points, and a utility function with concave and convex segments. It draws on Allais for its treatment of probabilities.

I am currently working with Meir Statman, a noted behavioral finance aficionado on procedures that speak to investors in terms of the "mental accounts" of behavioral finance, yet produce portfolios that are nearly mean-variance efficient.

Markowitz (1952a) Versus Markowitz (1959)

Markowitz (1952a) differs from Markowitz (1959) in various ways as outlined in the following paragraphs.

Markowitz (1952a) proposed mean-variance both as a maxim for recommended behavior and as a hypothesis concerning actual investor behavior. In Markowitz (1959), no mention is made of mean variance as a hypothesis about actual behavior.

Neither Markowitz (1952a) nor Markowitz (1959) assumed that mean-variance was to be applied to a "single-period world." Both considered mean-variance analysis as a single-period analysis within a many-period world. Markowitz (1952a) gave two proposals for the relationship between the single-period analysis and the many-period world. The first proposal considers the means, variances, and covariances of portfolio analysis as applicable to the present value of future dividends. The alternate proposal assumes that probability distributions of returns are in a steady state, and considers the means, variances, and covariances of the analysis to be the parameters of this steady state. On the other hand, Markowitz (1959), Chaps. 11 and 13, considers the single-period analysis to be applicable to the quadratic approximation (see below) to the derived utility function of dynamic programming.[1]

Tobin (1958) notes that mean-variance analysis is implied if the user has a quadratic utility function or if probability distributions are Gaussian. Neither Markowitz (1952a) nor Markowitz (1959) assumes Gaussian probability distributions. Markowitz (1952a) gives no justification for mean-variance analysis other than that variance (or, equivalently, standard deviation) is a commonly used measure of dispersion. Markowitz (1959), Chaps. 6 and 13, assumes that the investor should act to maximize expected utility, and proposes mean-variance analysis as an approximation to the maximization of expected utility. This view is spelled out more completely in Young and Trent (1964) for mean-variance approximations to the geometric mean or expected logarithm, and in Levy and Markowitz (1979) for various utility functions. More recent studies include Dexter et al. (1980), Pulley (1981, 1983), Kallberg and Ziemba (1981, 1983, 1984), Kroll et al. (1984), Simaan (1993), and Hlawitschka (1994).

The conclusions of the above studies are generally supportive of mean-variance approximations, including Hlawitschka's conclusion that it is applicable to *portfolios* of puts and calls. 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 the portfolio's bankruptcy.

Markowitz (1956) Versus Markowitz (1959)

Markowitz (1956) and Markowitz (1959), Chap. 8 and Appendix A, present the critical line algorithm for tracing out the entire mean-variance efficient set. Markowitz (1959), Appendix A, shows that the algorithm works even if the covariance matrix is singular.

[1] Please check if the sentence "On the other hand...." conveys the intended meaning.

I am understandably delighted with the results of Niedermayer and Niedermayer (2008), which appear in the present volume, an indication of how amazingly fast their variant of the critical line algorithm is when compared to alternate methods for tracing out the entire efficient frontier.

Sharpe (1964) Versus Markowitz (2005)

Major results from Sharpe (1964) and Lintner (1965) are that, given the assumptions of CAPM, the market portfolio is a mean-variance efficient portfolio, and there is a linear relationship between the excess return (expected return minus the risk-free rate) of a security and its beta (regression against market return). Markowitz (2005) points out that when investors cannot borrow all they want at the risk-free rate or, alternatively, cannot short and use the proceeds to buy long (which, in fact, is not a realistic model of short sales), then typically the market is not an efficient portfolio, and typically there is no linear relationship between excess return and beta.

Sharpe (1964) and Mossin (1966) Versus Markowitz (2008)

The assumptions of CAPM imply a linear relationship between excess return and the beta of a security, defined as its regression against the return on the market portfolio. This was interpreted as the investor being paid to bear risk. In fact, as explained in Markowitz (2008), causation goes in the other direction. The CAPM assumptions (as formalized by Mossin) imply that securities with higher expected returns per share have their prices bid up so they become a larger fraction of the market portfolio. Specifically, their prices are bid up to the point where the excess return per dollar invested is proportional to the regression of each security against the market portfolio, and where the market portfolio (of course) includes the security itself.

Markowitz and van Dijk (2003) and Kritzman, Myrgren and Page (2007)

As explained in Chap. 13 of Markowitz (1959), mean-variance analysis assumes perfectly liquid assets. This is not realistic, especially in the face of unrealized capital gains. Markowitz (1959) conjectures various procedures that can be used in the case of various kinds of illiquidities. But these are unevaluated conjectures only. Markowitz and van Dijk present a heuristic for dealing with problems with illiquid assets and changing probability distributions. They show that their "quadratic surrogate" heuristic works remarkably well for a problem that is small enough for its optimum solution to be computed. Kritzman, Myrgren and Page (2007) test the heuristic on some larger, practical problems and find that it works remarkably well on these also.

Harry Markowitz

References

Dexter, A.S., J.N.W. Yu and W.T. Ziemba (1980) Portfolio Selection in a Lognormal Market When the Investor Has a Power Utility Function: Computational Results. In: M.A.H. Dempster (ed.) Stochastic Programming. Academic, New York, pp. 507–23.

Friedman, M. and L.J. Savage (1948) The Utility Analysis of Choices Involving Risk. Journal of Political Economy 56:279–304.

Grauer, R.R. (1986) Normality, Solvency, and Portfolio Choice. Journal of Financial and Quantitative Analysis 21(3):265–78.

Hlawitschka, W. (1994) The Empirical Nature of Taylor-Series Approximations to Expected Utility. The American Economic Review 84(3):713–9.

Kahneman, D. and A. Tversky (1979) Prospect Theory: An Analysis of Decision Under Risk. Econometrica 47(2):263–91.

Kallberg, J.G. and W.T. Ziemba (1981) Remarks on Optimal Portfolio Selection. In: G. Bamberg and O. Optiz (eds.) Methods of Operations Research. Oelgeschlager, Gunn & Hain, Cambridge, MA, pp. 507–20.

Kallberg, J.G. and W.T. Ziemba (1983) Comparison of Alternative Utility Functions in Portfolio Selection Problems. Management Science 29(11):1257–76.

Kallberg, J.G. and W.T. Ziemba (1984) Mis-specifications in Portfolio Selection Problems. In: G. Bamberg and K. Spremann (eds.) Risk and Capital. Springer, New York, pp. 74–87.

Kroll, Y., H. Levy and H.M. Markowitz (1984) Mean Variance Versus Direct Utility Maximization. Journal of Finance 39(1):47–61.

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 47:13–37.

Markowitz, H.M. (1952a) Portfolio Selection. The Journal of Finance 7(1):77–91.

Markowitz, H.M. (1952b) The Utility of Wealth. The Journal of Political Economy 60:152–8.

Markowitz, H.M. (1956) The Optimization of a Quadratic Function Subject to Linear Constraints. Naval Research Logistics Quarterly 3:111–33.

Markowitz, H.M. (1959) Portfolio Selection: Efficient Diversification of Investments. Wiley, Yale University Press, 1970, 2nd edn. Basil Blackwell, 1991.

Markowitz, H.M. (2005) Market Efficiency: A Theoretical Distinction and So What? Financial Analysts Journal 61(5):17–30.

Markowitz, H.M. (2008) CAPM Investors Do Not Get Paid for Bearing Risk: A Linear Relation Does Not Imply Payment for Risk. The Journal of Portfolio Management 34(2):91–4

Mosteller, F. and P. Nogee (1951) An Experimental Measurement of Utility. Journal of Political Economy LIX(5):371–404.

Pulley, L.M. (1981) A General Mean-Variance Approximation to Expected Utility for Short Holding Periods. Journal of Financial and Quantitative Analysis 6:361–73.

Pulley, L.M. (1983) Mean-variance Approximations to Expected Logarithmic Utility. Operations Research 31(4):685–96.

Sharpe, W.F. (1964) Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk. The Journal of Finance 9(3):425–42.

Shefrin, H. (ed.) (2001) Behavioral Finance, Vol. 3. Edward Elgar, Northampton, MA.

Simaan, Y. (1993) What Is the Opportunity Cost of Mean-Variance Investment Strategies? Management Science 39(5):578–87.

Tobin, J. (1958) Liquidity Preference as Behavior Towards Risk. Review of Economic Studies 25(1):65–86.

Thaler, R. and E. Johnson (1990) Gambling with the House Money and Trying to Break Even: The Effects of Prior Outcomes on Risky Choice. Management Science 36(6):643–60.

Preface

Harry Markowitz and Professor Martin J. Gruber shared a keynote address at The International Symposium on Forecasting (ISF) in Nice, France, in June 2008. Harry is President of the Harry Markowitz Company, San Diego, CA and Professor Emeritus, Economics and Finance at Baruch College, City University of New York. Professor Martin J. Gruber is the Nomura Professor of Finance at The Stern School of Business, New York University. This volume was created to honor Harry for his address. We had two sessions at the ISF on applications of Markowitz analysis and portfolio construction, and eight papers in this volume were presented in Nice. Harry created much of the portfolio theory with his seminal monograph, *Portfolio Selection* (Yale University Press, 1959). Dr. Markowitz is perhaps best known as a co-recipient of the 1991 Nobel Prize in Economic Sciences. Most of the research on Wall Street has developed from this volume. When Harry agreed to deliver the keynote address 3 years ago, I wrote to 20 of Harry's friends, co-authors, and persons often cited in his papers regarding contributing to this volume. This volume includes 15 contributions from the 20 persons that were invited. Most people responded almost immediately, and several persons said "I love the man," "I have made a successful career of implementing Harry's models," or "I am glad that someone is doing it." I created this volume for several reasons: (1) to honor Mr. Markowitz for his lasting contribution to portfolio construction and testing; (2) to have a single source where someone who knew nothing of economics or investments could read the introductory chapters and have an idea as to how numbers are crunched in finance and why; (3) to have this volume produced while Harry and the rest of us could appreciate it; Harry is over 65 years old, as are several of the contributors. It is my hope and expectation that Harry will live at least another 20–25 years. This volume is composed of "Wall Street" and academic contributors, Harry having influenced both worlds. I must thank Professor Marty Gruber, who has been a great friend for many years, and who was kind enough to bring Ellie, his lovely wife, to Nice. Marty contributed a chapter based on several of his published articles, and was very helpful in reading papers and advising me on different aspects of the volume. Marty and Ned Elton edited an excellent volume for Harry in 1979, which is referenced throughout this volume. This volume updates much of the portfolio construction management and measurement analysis of the 1980–2008 period.

I would like to thank Robert B. (Bob) Gillam (Bob) and Robert A. (Rob) Gillam of McKinley Capital Management, LLC, for sponsoring the Markowitz address. McKinley Capital Management is a firm that implements Markowitz portfolio construction and management. As a graduate student at UCLA, Bob was influenced by Dr. Markowitz on what became known as Modern Portfolio Theory (MPT). Rob has continued his Dad's work in portfolio construction and has shown great leadership in helping McKinley grow and develop into a Global Growth specialist.

The majority of the papers in this volume have been peer-reviewed. Several of the chapter authors reviewed other manuscripts. I would like to acknowledge other peer reviewers involved with this volume.

Professor Stephen J. Brown, New York University
Ms. Kathleen DeRose, Hagin Investment Management
Dr. James L Farrell, Jr., Chairman, The Institute for Quantitative Research in Finance (The Q-Group), and Ned Davis Research, Inc.
Dr. Steven Greiner, Allegiant Asset Management Group
Dr. Robert Hagin, Hagin Investment Management
Mr. Gil Hammer, The Q-Group
Professor Mustafa Gultekin, The University of North Carolina
Professor Javier Peña, Carnegie Mellon University
Dr. Bernd Scherer, Morgan Stanley
Dr. David Stein, Parametric Portfolio Associates
Professor Sheridan Titman, The University of Texas, Austin
Professor Arnold Zellner, The University of Chicago
Dr. Victor Zarnowitz, The Conference Board

I serve as Director of Quantitative Research at McKinley Capital Management in Anchorage, AK. I co-managed two Japanese equity funds, Fund Academy in Tokyo, and the Japan Equity Find (JEQ), a closed-end country fund traded on the New York Stock Exchange, with Dr. Markowitz at Daiwa Securities. Harry and our group at Daiwa Securities co-authored several papers that are referenced in the introductory chapters. This book was edited and my three introductory chapters written on Saturday afternoons and Sunday mornings. My chapter on the 130/30 model was written as a McKinley Capital contribution and was presented in Nice, a spectacular city. I serve on the editorial board of the *International Journal of Forecasting* (IJF), the journal associated with the annual ISF. I suggested Nice as a possible venue for the address and could not have been more pleased with the city and the weather (sunny and 90).

I have dedicated the volume to three persons: Julie, my wife, who continues to demonstrate the patience of Job with me, and Bob and Rob Gillam, my bosses. If I can keep these three persons happy (at least satisfied), then my life is successful, or at least worth living. I am a big believer that life is pass/fail. The Wall Street experiences post June-2008 re-enforce my thoughts on that aspect of life. Rene Hollinger typed the first three chapters, as she did for my Springer volume with Eli Schwartz. Nick Philipson of Springer was very supportive of the project. Julie and I have three children, Richard, Katherine, and Stephanie, who continue to be a source of pleasure and pride. To Julie, Bob Gillam, Rob Gillam and Harry Markowitz, I say, thank you.

Anchorage, AK *John Guerard*

Contents

Part I
Markowitz for the Masses: Portfolio Construction Techniques

Chapter 1
Markowitz for the Masses: The Risk and Return of Equity and Portfolio Construction Techniques

John B. Guerard, Jr.

1.1 Introduction

Markowitz analysis seeks to maximize return for a given level of risk, or minimize risk for a given level of return. Prior to Harry Markowitz, investments were often associated with returns and not risk. John Burr Williams, in his seminal *The Theory of Investment Value* (1938), stated in his preface Investment Value, defined as the present value of future dividends or of future coupons and principal, is of practical importance to every investor because it is the critical value above which he cannot go in buying or holding without added risk. If a man buys a security below its investment value he need never lose, even if its price should fall at once, because he can still hold for income and get a return above normal on his cost price. But if he buys it above its investment value, his only hope of avoiding a loss is to sell to someone else who must in turn take the loss in the form of insufficient income (p. viii). Mr. Williams put forth a theory of investment value that influenced Mr. Markowitz profoundly. Let us review the work of Mr. Williams. For stocks, Williams calculated the present value of future dividends:

$$V_0 = \sum_{t=1}^{\infty} d_t \left(\frac{1}{1+i} \right)^t \tag{1.1}$$

where V_0 = Investment value at start, d_t = dividend in year t, and i = interest sought by investor.

$$V_0 = d_1 \left(\frac{1}{1+i} \right) + d_2 \left(\frac{1}{1+i} \right)^2 + d_3 \left(\frac{1}{1+i} \right)^3 + \cdots \tag{1.2}$$

J.B. Guerard, Jr. (✉)
McKinley Capital Management, LLC, Anchorage, AK, USA
e-mail: jguerard@mckinleycapital.com

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_1,
© Springer Science+Business Media, LLC 2010

The reader in 2009 will note that (1.2) is well known in financial analysis and is an early form of the Gordon (1962) model.[1] Williams found the present value of dividends, not earnings, because the earnings not paid out as dividends should be successfully reinvested for stockholder benefit such that they produce future dividends, or there is money lost (p. 57). Earnings are a means to dividends. Williams, quoting the old farmer to his son (p. 58):

> A cow for her milk,
> A hen for her eggs,
> And a stock, by heck,
> For her dividends.
> An orchard for fruit,
> Bees for their honey,
> And stocks, besides,
> For their dividends.

Williams, in discussing dividends and earnings, is very clear that in the long run, high earnings must be produced for high dividends to be paid.

If a firm maintains a constant (real) dividend policy, then (1.1) may be rewritten as

$$V_0 = \frac{d}{i} \tag{1.4}$$

where d = real dividend per year.

A stock is worth the present value of the dividends to be paid upon it, no more and no less. Williams (p. 80) declared that present earnings, financial condition, and (economic) outlook only assist buyers and sellers in estimating future dividends. Why are earnings retained by corporations? Williams holds that a certain percentage of earnings must be reinvested in the firm to maintain its competitive position and therefore maintain its dividend-paying power, not, in the case of mature companies, increase its dividend-paying power. Williams (p. 82) briefly discusses the price-earnings multiple, a variable that we will discuss and analyze repeatedly in this monograph, and notes that the "10 times earnings" multiple may not be relevant for a non-growing firm with low reinvestment needs and low depreciation charges (p. 82).

A further complication in the Williams theory of investment value is the valuation of net quick assets. A firm with cash holdings in excess of what is necessary to maintain its earnings power can pay an extra, non-recurring dividend. If the company has

---

[1] The Gordon model is written:

$$\begin{aligned} P_0 &= \int_0^\infty D_t^{\bar{e}^{kt}}\,dt \\ Dt &= Y_0(1-b)e^{rbt} \end{aligned} \tag{1.3}$$

where $D_0 = (1-b)Y_0$
$g = br$
b = retention rate of income, Y_0
$(1-b)$ = dividend pay-out rate.

receivables and inventories (or other quick or current assets) that are excessive, then these assets can be liquidated and an extra dividend paid:

$$V_0 = \frac{d}{i} + Q \tag{1.5}$$

where Q = excess of net quick assets per share.

Williams, in Chap. VII, discusses the valuation of "Stocks with Growth Expected" and examines the hypothetical case of forever-increasing dividends. Dividends are assumed to grow at the rate, g. Thus, $d_1 = d_0(1+g)$ and $d_2 = d_0(1+g)^2$.

$$d_t = d_0 u^t \tag{1.6}$$

$$V_0 = \sum_{t=1}^{\infty} d_t u^t V_t = d_0 \sum_{t=1}^{\infty} w^t \tag{1.7}$$

where $V = 1/(1+i)$

$$w = uv$$
$$d_0 \sum_{t=1}^{\infty} w^t = d_0 w(1 + w + w^2 + w^3 + \cdots)$$

If $g < i$, then $(1+g)/(1+i) < 1$ and $w = uv < 1$, and V_0 is finite. If $g \geq i$, then $(1+g)/(1+i) \geq 1$ and V_0 is infinite. Thus, one can price a stock if and only if the growth of its dividend-paying power is less than the interest rate used to discount future dividends (p. 89).[2] Williams used several firms to illustrate his theory of investment value as of November 1936. In the case of General Motors, GM, Chap. 21, Williams calculates 11-year average earnings per share and dividend per share of \$3.52 and \$2.90, respectively. Williams adjusts the average dividend retribution rate of 78.1 to 83.33% because the earlier portion of the 11-year period had only a 61.7% dividend payout ratio due to an expansion policy (in 1925–1929); the dividend payout ratio for 1930–1936 was approximately 94.6% and the GM share of the automobile share of USA and Canadian sales was rising from 33.4 to 44.1% (p. 398). Williams estimated from the term structure of interest rates as of June 15, 1937, that the proper interest rate over the next 5 years would be approximately 4.75% (Table 5, Chap. 20, p. 352). Thus, if the normal dividend of \$2.90 was capitalized at 4.75%, then the GM investment value as of June 1937 was \$61. By November 1936, general interest rates had fallen relative to June, as Williams reported from the prices of long-term government bonds, to about 4.50%, and the proper investment value of General Motors was \$64.50. Williams stated that despite

[2] On pp. 89–93, Williams uses an S-curve of dividend growth to establish the stock price of a stock that has dividends increasing rapidly, then slowly. The reader recognizes this analysis is reasonably similar to the "super normal" growth pricing of stocks.

the GM price of $77 at the time of the November Automobile Show the price "could not be called grossly extravagant as full markets go." The reader must laugh to him (her) self at the reading of November 1936 being considered a bull market compared with 2008, when economic history considers October 1929 to 1941 to be one of depression if not depression and recession.

Although Markowitz acknowledges Williams in his initial thoughts on common stock, one must consider other influential financial writers of the Williams era. Benjamin Graham and David Dodd, in their classical *Security Analysis* (1934), are presently considered by many, including Warren Buffet, the preeminent financial contributors to the theory of stock valuation. In Chap. 17, "The Theory of Common-Stock Investment," Graham and Dodd discussed their explanation for the departure of the public from rational common valuation during the 1927–1929 period. Graham and Dodd attributed much of the valuation departures to the instability of intangibles and the dominant importance of intangibles (p. 301). Investment in the pre-war (World War I) was confined to common stocks with stable dividends and fairly stable earnings, which would lead to stable market levels (p. 303). Graham and Dodd hold that the function of (security) analysis is to search for weakness in pre-war conditions such as improperly stated earnings, a poor current condition of its balance sheet, or debt growing too rapidly (p. 303). Moreover, new competition, deteriorating management, and/or market share must condemn the common stock from the possible valuation of a "cautious investor." Graham and Dodd attributed speculation on future prospects or expectations. They taught their students at Columbia University that buying common stock could be viewed as taking a share in the business (p. 305). In the new-era theory of common-stock valuation, (1) dividend rate had little bearing upon value; (2) there was no relationship between earning power and asset value; and (3) past earnings were significant only to the extent that they indicated what changes in earnings might occur in the future (p. 307). Graham and Dodd chastised 1927–1929 investors for valuation analysis that emphasized: (1) "the value of a common stock depends on what it can earn in the future; (2) good common stocks will prove sound and profitable investments; and (3) good common stocks are those which have shown a rising trend of earnings" (p. 309). The new-era investment theory was held to be the equivalent of pre-war speculation. Graham and Dodd attributed much of the speculation to the work of Edgar Smith and his *Common Stocks as Long-Term Investments* (1924) text. Graham and Dodd held that Smith postulated that stocks increased in value more than was justified as a result of reinvestment earnings capitalization. That is, if a company earned 9%, paid dividends of 6%, and added 3% to surplus (retained earnings), then good management should lead the stock value to increase with its book value, with the 3% compounded. Graham and Dodd assaulted the new-era investment theory for paying 20–40 times earnings. Smith, with his reinvestment of surplus earnings theory, built up asset values and thus created the growth of common-stock values (Graham and Dodd, p. 313). Average earnings, as advocated by Williams previously noted, ceased to be a dependable measure of future earnings, and there is a danger of projecting trends into the future. Graham and Dodd, in Chap. 23, "A Proposed Canon of Common-Stock Investment," formulated their canon of common-stock investment in which the (1) diversification of risk is used to yield a favorable average of results, and hence investment is a group operation; (2) individual issues are selected by qualitative and quantitative tests; and (3) a greater effort is made to determine the future outlook (p. 317). Thus, the purchase price of stock must have a rational basis and the reasonableness criterion is dependent upon past records of performance, such that "strongly entrenched companies are almost certain to fare better than those with poor past earnings and unsatisfactory balance-sheet positions" (p. 319). Satisfactory current and future dividends should be required for group purchases of common stocks. Purchases of common stock must involve "safety," with the intention and reasonable expectation of securing principal (p. 321). The earnings ratio, or earnings rate, is the ratio of annual earnings per share to the market price (p. 333). Graham and Dodd stated that prior to 1927–1929, "ten times earnings" was the accepted standard of measurement (p. 451). Graham and Dodd held that the maximum appraisal for common-stock investment should use average earnings of no less than 5 years, preferably 7 to 10 years, with a preference given to common stocks with current earnings above the average, and that "about sixteen times average earnings is as high a price as can be paid in an investment purchase of a common stock" (p. 453). Thus, an investor who "pays more than 16 times earnings will lose considerable money in the long run" (p. 453). Graham and Dodd emphasized the need for a strong working-capital position (where current assets exceed current liabilities).

In Chap. 18, Graham and Dodd discussed the current-asset rule, which held that the current-asset value is an approximation of liquidating value. It is illogical for firms to sell persistently below their liquidating value (p. 495). Current assets, less all liabilities, are the Graham and Dodd measure of current-asset value (p. 496). One sees that the net current asset value rule is an early form of the economics underlying a leveraged payout.

By 1950, Benjamin Graham had evolved his thinking into *The Intelligent Investor* and held that the factors affecting the capitalization rate were: (1) general long-term prospects that can cause differences in industry and individual company earnings ratios or multiples; (2) management; (3) financial strength and capital structure, as noted previously with surplus cash; (4) dividend record, with at least 20 years of dividends; and (5) the current dividend yield (pp. 154–157). Graham (1950) discussed a standard dividend policy in which two-thirds of earnings are paid out as dividends (on average, p. 157). Graham advocated a value of:

$$\text{Value} = \text{Current (Normal)Earnings} \times (8.5 \text{ plus twice the expected annual growth rate})$$

The growth rate is the one expected over the next 7 to 10 years. The 8.5 multiple times two times the growth rate translates into 17 times current earnings (p. 158). Graham elaborated further in valuation case studies. Companies should produce high rate of return on invested capital to accompany high annual growth rates in earnings per share (p. 177). Firms should not have instability. Rather, with the stability criteria, no preferred companies with no earnings decline and few companies with earnings decline exceeding the decline of earnings for the Dow Jones Industrial Average (p. 178). Firms should have earnings growth exceeding the Dow, hence a tendency toward lower price-earnings multiples. Companies should exceed a standard ratio of two times current assets relative to current liabilities.

Finally, dividends should have been continued without interruption. Firms should have adequate size, with annual sales not less than $100 million for an industrial company or $50 million for a public utility (p. 184). No current price should exceed 15 times average earnings of the past 3 years (p. 185). The current price should not exceed 15 times book value and the produce of the price-earnings multiple and the price-book multiple should not exceed 22.5 (p. 185). The stock portfolio should have an earnings/price ratio at least as high as the current high-grade bond rate (p. 186).[3]

Individual investors must be compensated for bearing risk. It seems intuitive to the reader that there should be a direct linkage between the risk of a security and its rate of return. We are interested in securing the maximum return for a given level of risk, or the minimum risk for a given level of return. The concept of such risk-return analysis is the efficient frontier of Harry Markowitz (1952, 1959). If an investor can invest in a government security, which is backed by the taxing power of the Federal Government, then that government security is relatively risk-free. The 90-day Treasury bill rate is used as the basic risk-free rate. Supposedly, the taxing power of the Federal government eliminates default risk of government debt issues. A liquidity premium is paid for longer-term maturities, due to the increasing level of interest rate risk. Investors are paid interest payments, as determined by the bond's coupon rate, and may earn market price appreciation on longer bonds if market rates fall or losses vice versa. During the period from 1926 to 2003, Treasury bills returned 3.69%, longer-term government bonds earned 5.28%, corporate bonds yielded 5.99%, and corporate stocks, as measured by the stock of the S&P 500 index, earned 11.84% annually. Small stocks averaged a 16.22% return, annually, over the corresponding period. The annualized standard deviations are 1.00, 19.48, and 29.66%, respectively, for Treasury bills, stocks (S&P), and small stocks. The risk-return trade off has been relevant for the 1926–2003 period.

Markowitz (1991) acknowledges a great intellectual debt to Professor James Tobin of Yale University. Tobin, in his 1958 seminal article on liquidity preference,

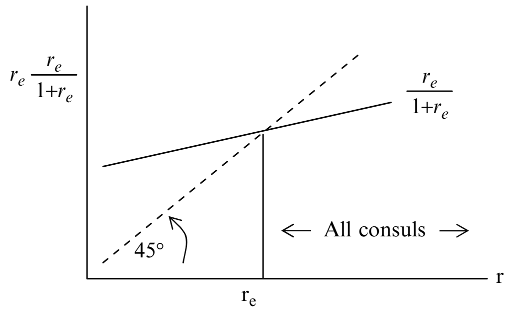


Fig. 1.1 Investment process

[3] In 1970, in the fifth edition of *The Intelligent Investor*, Graham found that only five companies of the Dow Jones Industrial Average firms satisfied his criteria.

analyzed the risk-return trade-off of two assets, cash, and a monetary asset, consuls. A consul is a (perpetual) bond that pays a current yield of r (per year). The expected yield on cash is zero. The investor expectations of the rate on consuls changes with market conditions. If the investor expects the rate on consuls to be r_e at the end of the year, then the capital gain (or loss) on the consul is given by g, where

$$g = \frac{r}{r_e} - 1 \tag{1.8}$$

An investor divides his (or her) balance between A_1 of cash and A_2 in consuls. If the current rate plus the expected capital gain exceeds zero, then the investor invests all his resources in consuls. If $r + g$ is expected to be less than zero, then the investor invests exclusively in cash. Tobin expressed a critical level of the current rate, r_c, where

$$r_c = \frac{r_e}{1 + r_e} \tag{1.9}$$

If the current rate exceeds r_c, then all funds are invested in consuls; however, if r is less than r_c, then all funds are invested in cash.

An investor is fully invested such that $A_1 + A_2 = 1$. The return on the portfolio, R, is

$$R = A_2(r + g) \tag{1.10}$$

where $0 \leq A_2 \leq 1$.

If g is a random variable with an expected value of zero, then

$$E(R) = \mu_r = A_2 r \tag{1.11}$$

The mean value of the portfolio return is μ_r and the standard deviation is a measure of portfolio dispersion, σ_R. A high standard deviation portfolio offers the investor a chance of large returns, as well as an equal chance of large losses. The standard deviation of the portfolio depends exclusively upon the standard deviation of consuls, and the percentage assets invested in consuls.

$$\sigma_R = A_2 \sigma_g \quad \text{where} \quad 0 \leqslant A_2 \leqslant 1 \tag{1.12}$$

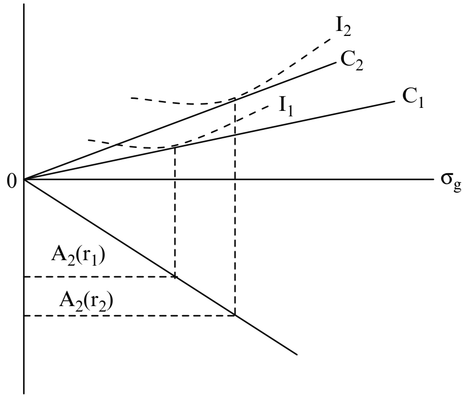
The ratio of expected return relative to risk can be written as

$$\mu_R = \frac{r}{\sigma_g} \sigma_R, \quad 0 \leqslant \sigma_R \leqslant \sigma_g \tag{1.13}$$

Tobin discusses the indifference of pairs (μ_R, σ_R) of risk and return that lie on the utility indifference curve I_1. Points on the I_2 indifference curve dominate points of the I_1 indifference curve. For a given level of risk, an investor prefers more returns to lesser returns. Risk-averters will accept more risk only if they are compensated with higher expected returns.

There is a risk premium on stocks relative to the risk-free rate, as reflected in the 90-day Treasury bill yield, R_R. If investors can borrow and lend at the risk-free asset

Fig. 1.2 The critical line



(with unlimited funds), then one can invest x percent of the portfolio in risky assets with expected return, R_A, and $(1-x)$ percent of the portfolio in the risk-free asset. The expected return of the combined portfolio is

$$E(R_p) = (1-x)R_F + xE(R_A) \tag{1.14}$$

The risk of the combined portfolio, as measured by its standard deviation, is given by:

$$\sigma_p = \left[(1-x)^2\sigma_F^2 + x^2\sigma_A^2 + 2x(1-x)\sigma_A\sigma_F\rho_{FA}\right]^{1/2} \tag{1.15}$$

where ρ_{FA} = correlation coefficient between risky and risk-free assets.

The standard deviation of the risk-free rate, σ_F, is assumed to be zero.

$$\sigma_\upsilon = \left(x^2\sigma_A^2\right)^{1/2} = x\,\sigma_A \tag{1.16}$$

The percentage invested in the risk-free asset, X, can be found by:

$$\frac{\sigma_p}{\sigma_A} = x \tag{1.17}$$

The expected return of the combined portfolio, shown in (1.15), may be written using (1.17) as

$$E(R_p) = \left(1 - \frac{\sigma_p}{\sigma_A}\right) R_F + \frac{\sigma_p}{\sigma_A} \tag{1.18}$$

$$E(R_p) = R_F + \left(\frac{R_A - R_F}{\sigma_A}\right)\sigma_p \tag{1.19}$$

Graphically, one sees (1.19) as the risk-return trade-off between risky and risk-free assets.

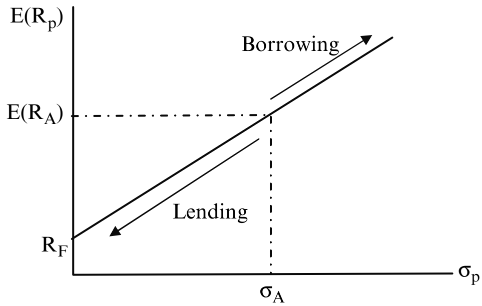


Fig. 1.3 Borrowing and risk-return trade-off analysis.

Let us examine the question of optimal weighting of risky and risk-free assets, given a two asset portfolio. In (1.14), we saw the expected return of a two asset portfolio,

$$E(R_\mathrm{p}) = (1 - x)R_\mathrm{F} + XE(R_\mathrm{A})$$

The standard deviation of the two asset portfolio was shown in (1.15),

$$\sigma_\mathrm{p} = \left[(1 - x)^2\sigma_\mathrm{F}^2 + x^2\sigma_\mathrm{A}^2 + 2(1 - x)(x)\sigma_\mathrm{A}\sigma_\mathrm{F}\rho_\mathrm{FA}\right]^{1/2}$$

We must take the derivative of (1.15) with respect to our decision variable, the percentage invested in the risky asset, and settle derivative to zero.

$$\frac{\partial\sigma_\mathrm{p}}{\partial x} = 1/2\left[\frac{2X\sigma_\mathrm{A}^2 - 2\sigma_\mathrm{F}^2 + 2X\sigma_\mathrm{A}^2 + 2\sigma_\mathrm{A}\sigma_\mathrm{F}\rho_\mathrm{FA} - X\sigma_\mathrm{A}\sigma_\mathrm{F}\rho_\mathrm{AF}}{X^2\sigma_\mathrm{A}^2 + (1 - X)^2\sigma_\mathrm{F}^2 + 2X(1 - X)\rho_\mathrm{FA}\sigma_\mathrm{A}\sigma_\mathrm{F}}\right]$$

$$X = \frac{\partial\sigma_\mathrm{p}}{\partial x} = 0 = \frac{\sigma_\mathrm{F}^2 - \rho_\mathrm{FA}\sigma_\mathrm{A}\sigma_\mathrm{F}}{\sigma_\mathrm{A}^2 + \sigma_\mathrm{F}^2 - 2\sigma_\mathrm{A}\sigma_\mathrm{F}\rho_\mathrm{FA}}$$

The Tobin presentation of the risk-return trade-off is quite relevant for common stocks and forms the basis of much of the Markowitz (1959) presentation. When one owns stocks, one is paid a dividend and earns stock price appreciation. That is, an investor buys stock when he or she expects its stock price to raise and compensates the investor for bearing the risk of the stock's price movements. Investors have become aware in recent years that not all price movements are in positive directions. One can calculate returns on stocks with several different methods.

1.2 Calculating Holding Period Returns

A very simple concept is the holding period return (HPR) calculation, introduced in Markowitz (1959, p. 14) in which one assumes that the stock was purchased at last (period's) year's price and the investor earns a dividend per share for the current year and a price, appreciation (depreciation) relative to last year's price.

$$\mathrm{HPR}_t = \frac{D_t + P_t - P_{t-1}}{P_{t-1}}$$

where D_t = current year's dividend, P_t = current year's stock price, P_{t-1} = last year's stock price and HPR_t = current year's holding period return.

The assumption of annual returns is arbitrary, but well-known in finance. Markowitz (1959) used annual returns in his Chap. 2 of *Portfolio Selection* to illustrate holding period returns, expected returns, and variability. Let us examine ten widely held stocks: GE, Dominion Resources, IBM, DuPont, BP, Ford, Pfizer, Conoco Philips, Royal Dutch Petroleum, and Johnson & Johnson, for 1997–2006 period. The pricing data is taken from the Standard & Poor's *Stock Guide*. The *S&P Stock Guide* presents high and low prices during the calendar year. An average price (AvgP) can be calculated by simply summing the high and low prices and dividing by two, see Guerard and Schwartz (2007). The average price calculations for the ten securities are shown in Table 1.1. Changes in annual average prices create much of the variability in annual HPRs. For example, in 1997, the average price of GE rose from \$20.75 to \$28.83 in 1998. The \$8.08 price appreciation leads to an annual return of 40.94%.

$$\mathrm{GE\,HPR}_{1998} = \frac{\$.36 + \$28.83 - \$20.75}{\$20.75} = .4094$$

The price movement of GE in 1998 was slightly higher than Dominion Resources' gain of \$5.31 on a \$38.07 base, producing a holding period return of 20.73%. Dominion Resources' price fell in 2002, falling \$9.35 on a \$62.56 investment, or a loss of some 13.99% in the annual holding period return. Dominion Resources is often referred to by its ticker symbol, D. The GE and D holding period returns (HPRs) were consistently positive and rather large during the 1997–2006 period. The reader sees that the HPRs of stocks are often far more dependent upon stock price movements than the dividends received by investors. The HPRs of GE ranged from 48.31% in 1999 to −21.15% in 2002. One can estimate an expected return for GE by calculating the mean value of the annual HPRs during the 1998–2006 period. The expected return for GE during the 1997–2006 period was 10.91%, with a standard deviation of 25.58%. If the annual HPRs of GE are normally distributed, that is, the returns fall within the normal "Bell" curve, then 67.6% of the annual observations of GE returns should fall within the −14.67 and 36.49% range (one standard deviations). The reader immediately sees how wide the one standard deviation range is for annual returns. We can calculate in a similar manner the expected returns for the remaining stocks.

One sees that the returns are extremely volatile for IBM, having a standard deviation exceeding 39%. It may be worthwhile to calculate a coefficient of variation (CV) in which the standard deviation is divided by the expected return. The calculation of the CV is shown in Table 1.2, and leads the investor to recognize that DD, more than IBM or GE, produced greater variation for a given level of expected returns.

Moreover, Ford produced a negative expected return and standard deviation during the 1997–2006 time period.

Table 1.1

	GE					D				
Year	P High	P Low	P Avg	Dividends	HPR	P High	P Low	P Avg	Dividends	HPR
2006	38.49	32.06	35.28	1.03	0.0371	84.44	68.72	76.58	0.69	0.0069
2005	37.34	32.67	35.01	0.91	0.0780	86.97	66.51	76.74	2.68	0.2253
2004	37.75	28.88	33.32	0.82	0.2780	68.85	60.78	64.82	2.60	0.1456
2003	32.42	21.00	26.71	0.77	−0.1309	65.95	51.74	58.85	2.58	0.1990
2002	41.84	21.40	31.62	0.73	−0.2115	67.06	35.40	51.23	2.58	−0.1399
2001	53.55	28.50	41.03	0.66	−0.1838	69.99	55.13	62.56	2.58	0.2679
2000	60.50	41.65	51.08	0.57	0.2221	67.94	34.81	51.38	2.58	0.2556
1999	53.17	31.35	42.26	0.49	0.4831	49.38	36.56	42.97	2.58	0.0501
1998	34.65	23.00	28.83	0.42	0.4094	48.94	37.81	43.38	2.58	0.2073
1997	25.52	15.98	20.75	0.36		42.88	33.25	38.07	2.58	
Mean					0.1091					0.1353
Std					0.2558					0.1365
CV					2.3460					1.0083

	IBM					DD				
Year	P High	P Low	P Avg	Dividends	HPR	P High	P Low	P Avg	Dividends	HPR
2006	97.88	72.73	85.31	1.10	0.0109	49.68	38.52	44.10	1.48	−0.0145
2005	99.10	71.85	85.48	0.78	−0.0980	54.90	37.60	46.25	1.46	0.0689
2004	100.43	90.82	95.63	0.70	0.1487	49.39	39.88	44.64	1.40	0.0883
2003	94.54	73.17	83.86	0.63	−0.0634	46.00	38.60	42.30	1.40	0.0304
2002	126.39	54.01	90.20	0.59	−0.1289	49.80	35.02	42.41	1.40	0.0618
2001	124.70	83.75	104.23	0.55	−0.0253	49.88	32.64	41.26	1.40	−0.2395
2000	134.94	80.06	107.50	0.51	−0.0184	74.00	38.19	56.10	1.40	−0.0819
1999	139.19	80.88	110.04	0.47	0.5479	75.19	50.06	62.63	1.40	−0.0594
1998	94.97	47.81	71.39	0.44	0.6227	84.44	51.69	68.07	1.37	0.1958
1997	56.75	31.78	44.27	0.39		69.75	46.38	58.07	1.23	
Mean					0.1107					0.0056
Std					0.2809					0.1241
CV					2.5375					22.3500

	BP					F				
Year	P High	P Low	P Avg	Dividends	HPR	P High	P Low	P Avg	Dividends	HPR
2006	76.85	63.52	70.19	2.30	0.1215	9.48	1.06	5.27	0.35	−0.4964
2005	72.66	56.60	64.63	2.09	0.2270	14.75	7.57	11.16	0.40	−0.2280
2004	62.10	46.65	54.38	1.66	0.3300	17.34	12.61	14.98	0.40	0.2861
2003	49.59	34.67	42.13	1.53	−0.0323	17.33	6.58	11.96	0.40	−0.0167
2002	53.98	36.25	45.12	1.41	−0.0447	18.23	6.90	12.57	0.40	−0.4378
2001	55.20	42.20	48.70	1.29	−0.0364	31.42	14.70	23.06	1.05	−0.3892
2000	60.63	43.13	51.88	1.22	0.0329	57.25	21.69	39.47	2.30	−0.2680
1999	62.63	40.19	51.41	1.52	0.2431	67.88	46.25	57.07	1.88	0.1397
1998	48.66	36.50	42.58	1.21	0.1095	65.94	37.50	51.72	2.18	0.3433
1997	46.50	32.44	39.47	1.12		50.25	30.00	40.13	1.65	
Mean					0.1056					−0.1186
Std					0.1375					0.3175
CV					1.3021					−2.6777

Table 1.1 (continued)

	PFE					COP				
Year	P High	P Low	P Avg	Dividends	HPR	P High	P Low	P Avg	Dividends	HPR
2006	28.60	22.16	25.38	0.96	0.0647	74.89	54.90	64.90	1.44	0.1753
2005	29.21	20.27	24.74	0.76	−0.1623	71.48	41.40	56.44	1.18	0.4820
2004	38.89	21.99	30.44	0.68	−0.0398	45.61	32.15	38.88	0.90	0.3300
2003	36.92	27.90	32.41	0.60	−0.0232	33.02	26.80	29.91	0.82	0.1367
2002	42.46	25.13	33.80	0.52	−0.1501	32.05	22.02	27.04	0.74	−0.0585
2001	46.75	34.00	40.38	0.44	0.0300	34.00	25.00	29.50	0.70	0.1403
2000	49.25	30.00	39.63	0.36	−0.0197	35.00	17.97	26.49	0.68	0.1445
1999	50.04	31.54	40.79	0.31	0.2329	28.63	18.84	23.74	0.68	0.0452
1998	42.98	23.69	33.34	0.25	0.6746	26.63	20.09	23.36	0.68	0.0727
1997	26.67	13.44	20.06	0.23		26.13	18.69	22.41	0.67	
Mean					0.0675					0.1631
Std					0.2562					0.1591
CV					3.7971					0.9751
	RDS.A					JNJ				
Year	P High	P Low	P Avg	Dividends	HPR	P High	P Low	P Avg	Dividends	HPR
2006	72.38	60.17	66.28	2.08	0.0861	69.41	56.65	63.03	1.46	−0.0059
2005	68.08	57.79	62.94	2.27	0.2590	69.99	59.76	64.88	1.28	0.1657
2004	57.79	45.79	51.79	1.59	0.1943	64.25	49.25	56.75	1.1	0.0800
2003	52.70	36.69	44.70	1.46	−0.0374	59.08	48.05	53.57	0.93	0.0158
2002	57.30	38.60	47.95	1.26	−0.0527	65.89	41.40	53.65	0.8	0.0758
2001	64.15	39.75	51.95	1.21	−0.0845	60.97	40.25	50.61	0.7	0.1928
2000	65.69	50.44	58.07	1.1	0.1066	52.97	33.06	43.02	0.62	−0.0508
1999	67.37	39.56	53.47	1.34	0.0947	53.44	38.50	45.97	0.55	0.2151
1998	60.38	39.75	50.07	1.37	0.0141	44.88	31.69	38.29	0.49	0.3378
1997	59.44	42.00	50.72	1.06		33.66	24.31	28.99		
Mean					0.0645					0.1140
Std					0.1155					0.1240
CV					1.7916					1.0872

Table 1.2 Description statistics of holding period returns

Security	Mean	Std	CV
GE	0.1091	0.2558	2.3460
D	0.1353	0.1365	1.0083
IBM	0.1107	0.2809	2.5375
DD	0.0056	0.1241	22.3500
BP	0.1056	0.1375	1.3021
F	−0.1186	0.3175	−2.6777
PFE	0.0675	0.2562	3.7971
COP	0.1631	0.1591	0.9751
RDS.A	0.0645	0.1156	1.7916
JNJ	0.1140	0.1240	1.0872

Note that the calculations of expected returns and standard deviations allow the investor to allocate scarce resources on the basis of historic returns and risk. An investor should not allocate resources to only one security, as was the case with many Enron or Worldcom stockholders. Remember your grandmother's expression, "Do not put all of your eggs in one basket." Clearly the standard deviation of return may be minimized by investing in several assets, particularly if these assets are somewhat uncorrelated. An investor does not benefit from investing in two stocks, as opposed to only one security, if both stocks move in parallel. That is, if stock A rises 10% and stock B rises 9.50%, then it is not evident that the investor has any benefits to a second stock investment, particularly if the stocks are highly correlated. However, if GE has an expected return of 10.91% and Dominion Resources has an expected return of 13.53%, then an investor can purchase an equal dollar amount of each stock and reduce risk, if the stocks are not perfectly correlated with each other. The correlation coefficient, as the reader remembers, is the covariance of two series, divided by the product of the respective standard deviations. The correlation coefficient allows an investor to understand the statistical nature of a covariance because the correlation coefficient is bonded between -1 and $+1$. Low correlations coefficients imply that the assets are not good substitutes for one another, and diversification is enhanced by using assets with lower correlations. The covariance between two series is calculated as the sum of the product of the differences between each series and its respective mean. If the covariance of GE and IBM is positive, then this implies that when GE's return is above its mean or expected value, then IBM's return is above its mean. The correlation coefficient of the GE and IBM return series is 0.837, which is near one. Thus, an investor might not want to have only GE and IBM in a two asset portfolio. The correlation coefficient of GE and Dominion Resources is only 0.050, which is the lowest set of correlations in the three assets, and an investor would want to use GE and Dominion Resources, as opposed to GE and IBM to build a two asset portfolio, if the investor wants to minimize risk. Let us illustrate the essence of Markowitz diversification and portfolio construction. The most important question is what portfolio weights will minimize risk? (Table 1.3).

The portfolio variance is given by the weighted asset variances and covariances.

$$\sigma_{\mathrm{p}}^2 = x_1^2\sigma_1^2 + x_2^2\sigma_2^2 + 2x_1(1-x_1)\sigma_{12}$$
$$\sigma_{12} = \rho_{12}\sigma_1\sigma_2 \tag{1.20}$$

where ρ_{12} = correlation coefficients of assets 1, 2 and σ_{12} = covariance of assets 1, 2.

Table 1.3 Correlation matrix of HPRs

	GE	DD	IBM
GE	1.0000	.1634	.8370
DD	0.1634	1.0000	.0500
IBM	0.8370	.0500	1.0000

1.3 Minimizing Risk

To minimize risk, one seeks to allocate resources to assets such that the change in risk goes to zero with the change in the percentage invested in the asset. That is, risk minimization implies a first partial derivative of zero, with respect to the change in the asset weight.

$$\begin{aligned}
\text{Let } x_2 &= 1 - x_1 \\
\sigma_p^2 &= x_1^2\sigma_1^2 + (1-x_1)^2\sigma_2^2 + 2x_1(1-x_1)\sigma_{12} \\
&= x_1^2\sigma_1^2 + (1-x_1)^2\sigma_2^2 + 2x_1(1-x_1)\rho_{12}\sigma_1\sigma_2 \\
&= x_1^2\sigma_1^2 + (1-x_1)(1-x_1)\sigma_2^2 + 2x_1\rho_{12}\sigma_1\sigma_2 - 2x_1^2\rho_{12}\sigma_1\sigma_2 \\
\sigma_p^2 &= x_1^2\sigma_1^2 + \left(1 - 2x_1 + x_1^2\right)\sigma_2^2 + 2x_1\rho_{12}\sigma_1\sigma_2 - 2x_1^2\rho_{12}\sigma_1\sigma_2 \\
\frac{\partial\sigma_p^2}{\partial x_1} &= 2x_1\sigma_1^2 - 2\sigma_2^2 + 2x_1\sigma_2^2 + 2\rho_{12}\sigma_1\sigma_2 - 4x_1\rho_{12}\sigma_1\sigma_2 = 0 \\
2\sigma_2^2 - 2p_{12}\sigma_1\sigma_2 &= 2x_1\sigma_1^2 + 2x_1\sigma_2^2 - 4x_1\rho_{12}\sigma_1\sigma_2 \\
(\sigma_1^2 + \sigma_2^2 - 2\rho_{12}\sigma_1\sigma_2)x_1 &= \sigma_2^2 - \rho_{12}\sigma_1\sigma_2 \\
x_1 &= \frac{\sigma_2(\sigma_2 - \rho_{12}\sigma_1)}{\sigma_1^2 + \sigma_2^2 - 2\rho_{12}\sigma_1\sigma_2}
\end{aligned} \tag{1.21}$$

Equation (1.21) shows the risk-minimizing weight (percentage invested) of asset one in the portfolio.

In an equally weighted portfolio, $x_1 = x_2 = 0.50$. Let $x_1 =$ weight of GE and $x_2 =$ weight of IBM. The portfolio expected return is a weighted combination of asset expected returns.

$$\begin{aligned}
E(R_p) &= x_1E(R_1) + x_2E(R_2) \\
&= 0.5(0.1091) + 0.5(0.1107) = 0.1099
\end{aligned} \tag{1.22}$$

$$\begin{aligned}
s_p^2 &= (0.5)^2\,(0.2558)^2 + (0.5)^2\,(0.2809)^2 + 2\,(0.5)\,(0.5)\,(0.837)\,(0.2558)\,(0.2809) \\
&= 0.0164 + 0.0197 + 0.0300 \\
&= 0.0661
\end{aligned}$$

$$\sigma_p = \sqrt{0.0661} = .2571$$

The expected return on an equally weighted portfolio of DD and D stock is 10.99%, and its corresponding standard deviation is 25.71%. The portfolio return should fall within the range of −14.72 and 36.62% approximately 67.6% of the time. This range corresponds to the expected return plus and minus one standard deviation of return.

If one created an optimally weighted portfolio of GE and IBM using (1.21), one can solve for the optimal weights.

$$\begin{aligned}
x_1 &= x_{\mathrm{GE}} = \frac{\sigma_2(\sigma_2 - \sigma_1\rho_{12})}{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho_{12}} \\
&= \frac{0.2809(0.2809 - 0.2558(0.837))}{(0.2558)^2 + (0.2809)^2 - 2(0.837)(0.2558)(0.2809)} \\
\mathrm{GE} &= \frac{0.2809\ (0.0668)}{0.0654 + 0.0789 - 0.1203} \\
X_{\mathrm{GE}} &= \frac{0.0188}{0.0240} = \ 0.783 \\
x_{\mathrm{IBM}} &= 0.217
\end{aligned}$$

$$\begin{aligned}
E(R_{\mathrm{p}}) &= 0.783\,(0.1091) + 0.217(0.1107) \\
&= 0.0854 + 0.0240 = 0.1094 \\
\sigma_{\mathrm{p}}^2 &= (0.783)^2\,(0.2558)^2 + (0.217)^2\,(0.2809)^2 + 2\,(0.783)\,(0.2127)\,(0.2558) \\
&\quad (0.2809)\,(0.837) \\
&= 0.0401 + 0.0037 + (0.0204) \\
\sigma_{\mathrm{p}}^2 &= 0.0642 \\
\sigma_{\mathrm{p}} &= 0.2534
\end{aligned}$$

The optimally weighted portfolio return is 10.94%, and its standard deviation is only 25.34%. The portfolio variance is hardly reduced by using optimal portfolio weighting. Why? Because the GE and IBM holding period returns correlation coefficient is so near 1.000.

Let us calculate weights on a lesser-correlated example. Markowitz analysis or risk-minimization is based upon minimizing covariances or correlation coefficients.

	DD	IBM	D
$\bar{x}_i$	0.005	0.1107	0.1353
$\bar{\sigma}_i$	0.1241	0.2809	0.1365

Furthermore, let $\rho_{\mathrm{DD,\ IBM}} = 0.3147$, $\rho_{\mathrm{DD,\ D}} = -0.2108$, and $\rho_{\mathrm{D,\ IBM}} = 0.0500$.

$$\begin{aligned}
E(R_{\mathrm{p}}) &= X_{\mathrm{DD}}E(R_{\mathrm{DD}}) + X_{\mathrm{D}}E(R_{\mathrm{D}}) \\
\sigma_{\mathrm{p}}^2 &= x_1^2\sigma_{\mathrm{DD}}^2 + x_2^2\sigma_{\mathrm{D}}^2 + 2x_1x_2\sigma_{\mathrm{DD}}
\end{aligned}$$

In an equally weighted portfolio of DuPont, DD, and Dominion Resources, D,

$$x_{\mathrm{DD}} = 0.50 \quad \text{and} \quad x_{\mathrm{D}} = 0.50$$

The expected return on the portfolio may be calculated as

$$E\left(R_{\mathrm{p}}\right) = 0.5\,(0.0055) + 0.5\,(0.1353) = 0.0028 + 0.0677 = 0.0705$$

The expected portfolio variance may be calculated as

$$\begin{aligned}
\sigma_{\mathrm{p}}^2 &= (0.5)^2\,(0.1241)^2 + (0.5)^2\,(0.1365)^2 + 2\,(0.5)\,(0.5)\,(0.1241)\,(0.1365)\,(-0.218)\\
&= (0.5)^2\,(0.1241)^2 + (0.5)^2\,(0.1365)^2 + 2\,(0.5)\,(0.5)\,(0.1241)\,(0.1365)\,(-0.218)\\
&= 0.0039 + 0.0047 - 0.0018 = 0.0068\\
\sigma_{\mathrm{p}} &= 0.0825
\end{aligned}$$

How does the optimally weighted portfolio expected return and variance differ from the equally weighted portfolio statistics? Let us calculate the optimal portfolio weights.

$$\begin{aligned}
x_{\mathrm{DD}} &= \frac{\sigma(\sigma_{\mathrm{D}} - \rho_{\mathrm{DD,D}}\sigma_{\mathrm{DD}})}{\sigma_{\mathrm{DD}}^2 + \sigma_{\mathrm{D}}^2 - \rho_{\mathrm{DD,D}}\sigma_{\mathrm{DD}}\sigma_{\mathrm{D}}}\\
&= \frac{0.1365(0.1365 - (-0.2108)(0.1241))}{(0.1241)^2 + (0.1365)^2 - 2(-0.2108)(0.1241)(0.1365)}\\
&= \frac{0.1365(0.1365 - 0.0262)}{0.0154 + 0.0186 - (-0.0071)} = \frac{0.1365(0.1103)}{0.0411} = \frac{0.0151}{0.0411} = 0.364\\
x_{\mathrm{DD}} &= 0.364\\
x_{\mathrm{D}} &= 1 - 0.364 = 0.636
\end{aligned}$$

The investor should invest 36.4 percent of the portfolio in DuPont and 63.6 percent of the portfolio in Doninion Resources.

$$\begin{aligned}
E(R_{\mathrm{p}}) &= 0.364(0.0056) + 0.636(0.1353) = 0.0020 + 0.0861 = 0.0881\\
\sigma_{\mathrm{p}}^2 &= (0.364)^2(0.1241)^2 + (0.636)^2(0.1365)^2 + 2(0.364)(0.636)(-0.2108)\\
&\qquad (0.1241)\,(0.1365)\\
&= 0.0020 + 0.0075 - 0.0016 = 0.0079\\
\sigma &= \sqrt{0.0079} = 0.0889
\end{aligned}$$

The optimal portfolio has a higher return for a given level of risk than the equally weighted (EQ) portfolio. This can be shown in the portfolio coefficient of variation, CV.

$$\begin{aligned}
\mathrm{CV}_{\mathrm{p}}^{\mathrm{EQ}} \frac{\sigma_{\mathrm{p}}}{E(R_{\mathrm{p}})} &= \frac{0.0825}{0.0705} = 1.170\\
\mathrm{CV}_{\mathrm{p}} = \frac{\sigma_{\mathrm{p}}}{E(R_{\mathrm{p}})} &= \frac{0.0889}{0.0881} = 1.009
\end{aligned}$$

1.4 The Three Asset Case

Markowitz (1959) develops the well-known three asset case analysis in chapter 7.
Let us now examine a three asset portfolio construction process.

$$E(R_{\mathrm{p}}) \sum_{I=1}^{N} x_i E(R_1) \tag{1.23}$$

$$\sigma_{\mathrm{p}}^2 = \sum_{I=1}^{N} \sum_{I=1}^{N} x_i x_j \sigma_{ij} \tag{1.24}$$

$$E(R_{\mathrm{p}}) = x_1 E(R_1) + x_2 E(R_2) + x_3 E(R_3)$$

Let

$$\begin{aligned}
x_3 &= 1 - x_1 - x_2 \\
E(R_{\mathrm{p}}) &= x_1 E(R_1) + x_2 E(R_2) + (1 - x_1 - x_2) E(R_3) \\
\sigma_{\mathrm{p}}^2 &= x_1^2 \sigma_1^2 + x_2^2 \sigma_2^2 + \sigma_3^2 x_3^2 + 2x_1 x_2 \sigma_{12} + 2x_1 x_3 \sigma_{23} + 2x_2 x_3 \sigma_{23} \\
&= x_1^2 \sigma_1^2 + x_2^2 \sigma_2^2 + (1 - x_1 - x_2)^2 \sigma_3^2 + 2x_1 x_2 \sigma_{12} + 2x_1 (1 - x_1 - x_2) \sigma_{13} \\
&\quad + 2x_2 (1 - x_1 - x_2) \sigma_{23} \\
&= x_1^2 \sigma_1^2 + x_2^2 \sigma_2^2 + (1 - x_1 - x_2)(1 - x_1 - x_2) \sigma_3^2 + 2x_1 x_2 \sigma_{12} + 2x_1 \sigma_{13} \\
&\quad - 2x_1^2 \sigma_{13} - 2x_1 x_2 \sigma_{13} + 2x_2 \sigma_{23} - 2x_1 x_2 \sigma_{23} - 2x_2^2 \sigma_{23} \\
&= x_1^2 \sigma_1^2 + x_2^2 \sigma_2^2 + \left(1 - 2x_1 - 2x_2 + 2x_1 x_2 + x_1^2 + x_2^2\right) \sigma_3^2 + 2x_1 x_2 \sigma_{12} \\
&\quad + 2x_1 \sigma_{13} - 2x_1^2 \sigma_{13} - 2x_1 x_2 \sigma_{13} + 2x_2 \sigma_{23} - 2x_1 x_2 \sigma_{23} - 2x_2^2 \sigma_{23}
\end{aligned} \tag{1.25}$$

$$\begin{aligned}
\frac{\partial \sigma_{\mathrm{p}}^2}{\partial x_1} &= 2x_1 \left(\sigma_1^2 + \sigma_3^2 - 2\sigma_{13}\right) + x_2 \left(2\sigma_3^2 + 2\sigma_{12} - 2\sigma_{13} - 2\sigma_{23}\right) - 2\sigma_3^2 + 2\sigma_{13} = 0 \\
\frac{\partial \sigma_{\mathrm{p}}^2}{\partial x_2} &= 2x_2 \left(\sigma_2^2 + \sigma_3^2 - 2\sigma_{23}\right) + x_1 \left(2\sigma_3^2 + 2\sigma_{12} - 2\sigma_{13} - 2\sigma_{23}\right) - 2\sigma_3^2 + 2\sigma_{13} = 0
\end{aligned}$$

The equally weighted three asset case of DuPont, IBM, and Dominion Resources, using data in Table 1.2, can be written as

$$\begin{aligned}
E\left(R_{\mathrm{p}}\right) &= w_{\mathrm{DD}} E\left(R_{\mathrm{DD}}\right) + w_{\mathrm{IBM}} E\left(R_{\mathrm{IBM}}\right) + w_{\mathrm{D}} E\left(R_{\mathrm{D}}\right) \\
&= 0.333\,(0.0055) + 0.333\,(0.1107) + 0.333\,(0.1353) \\
&= 0.0018 + 0.0369 + 0.0451 = 0.0838 \\
\sigma_{\mathrm{p}}^2 &= (0.333)^2 (0.1241)^2 + (0.333)^2 (0.2809)^2 + (0.333)^2 (0.1365)^2 + 2\,(0.333) \\
&\quad (0.333)\,(0.3147)\,(0.1241)\,(0.2809) + 2\,(0.333)\,(0.333)\,(-0.2108)
\end{aligned}$$

$$
\begin{aligned}
&(0.1241)\,(0.1365) + 2\,(0.333)\,(0.333)\,(0.0500)\,(0.2809)\,(0.1365)\\
&= 0.0017 + 0.0087 + 0.0021 + 0.0024 + (-0.0008) + 0.0004\\
&= 0.0145\\
\sigma_{\mathrm{p}} &= 0.1196
\end{aligned}
$$

The equally weighted expected return is 8.38% and the corresponding expected portfolio variance is 11.96%.

The optimal weights for the DD, IBM, and D portfolio can be calculated as

$$
\begin{aligned}
\frac{\partial \sigma_{\mathrm{p}}^2}{\partial x_{\mathrm{DD}}} &= 2x_{\mathrm{DD}}(\sigma_{\mathrm{DD}}^2 + \sigma_{\mathrm{D}}^2 - 2\rho_{\mathrm{DD,D}}\sigma_{\mathrm{DD}}\sigma_{\mathrm{D}}) + x_{\mathrm{IBM}}\left(2\sigma_{\mathrm{DD}}^2 - 2\sigma_{\mathrm{DD,IBM}} - 2\sigma_{\mathrm{DD,D}}\right.\\
&\quad \left.-2\sigma_{\mathrm{IBM,D}}\right) - 2\sigma_{\mathrm{D}}^2 + 2\sigma_{\mathrm{DD,D}} = 0\\
&= 2x_{\mathrm{DD}}[(0.124)^2 + (0.1365)^2 - 2\,(-0.2108)\,(0.1241)\,(0.1365)]\\
&\quad + x_{\mathrm{IBM}}[2\,(0.1365)^2 - 2\,(0.3147)\,(0.1241)\,(0.2809) - 2\,(-0.2108)\\
&\quad (0.1241)\,(0.1365) - 2\,(0.0500)\,(0.2809)\,(0.1365)] - 2\,(0.1365)^2\\
&\quad + 2\,(-0.2108)\,(0.1241)\,(0.1365) = 0\\
&= 2x_{\mathrm{DD}}\,[0.0154 + 0.0186 + 0.0071] + x_{\mathrm{IBM}}\,[0.0373 - 0.0219 + 0.0071\\
&\quad - 0.0038] - 0.0373 + (-0.0071)\\
\frac{\partial \sigma_{\mathrm{p}}^2}{\partial x_{\mathrm{DD}}} &= 2x_{\mathrm{DD}}(0.0411) + x_{\mathrm{IBM}}(0.0187) - 0.0373 - 0.0071 = 0\\
&= 0.0822x_{\mathrm{DD}} + 0.0187x_{\mathrm{IBM}} - 0.0444 = 0\\
&\quad 0.0822x_{\mathrm{DD}} + 0.0187x_{\mathrm{IBM}} = 0.0444\\
\frac{\partial \sigma_{\mathrm{p}}^2}{\sigma x_{\mathrm{IBM}}} &= 2x_{\mathrm{IBM}}\left(\sigma_{\mathrm{IBM}}^2 + \sigma_{\mathrm{D}}^2 - 2\rho_{\mathrm{IBM,D}}\sigma_{\mathrm{IBM}}\sigma_{\mathrm{D}}\right)\\
&\quad + x_{\mathrm{DD}}\left(2\sigma_{\mathrm{D}}^2 - 2\sigma_{\mathrm{DD,IBM}}\sigma_{\mathrm{DD,D}} - 2\sigma_{\mathrm{DD,D}} - 2\sigma_{\mathrm{DD,D}}\sigma_{\mathrm{D}}\right)\\
&\quad \sigma_{\mathrm{IBM,D}} = 0\\
&= 2x_{\mathrm{IBM}}[(0.2809)^2 + (0.1365)^2 - 2\,(0.0500)\,(0.2809)\,(0.1365)\\
&\quad + x_{\mathrm{DD}}[2\,(0.1365)^2 + 2\,(0.3147)\,(0.2809)\,(0.1241)\\
&\quad - 2\,(-0.2108)\,(0.1241)\,(0.1365) - 2\,(0.0500)\,(0.2809)\,(0.1365)]\\
&\quad -2\,(0.1365)^2 + 2\,(0.0500)\,(0.2809)\,(0.1365) = 0\\
&= 2x_{\mathrm{IBM}}\,[0.0789 + 0.0186 - 0.0038] + x_{\mathrm{DD}}\,[0.0373 + 0.0219\\
&\quad + 0.0071 - 0.0038] - 0.0373 + 0.0038 = 0\\
&= 2x_{\mathrm{IBM}}\,(0.0937) + x_{\mathrm{DD}}\,(0.0702) - 0.0335 = 0\\
&= 0.1874x_{\mathrm{IBM}} + 0.0701x_{\mathrm{DD}} - 0.0335\\
&\quad 0.1874x_{\mathrm{IBM}} + 0.0701x_{\mathrm{DD}} = 0.0335\\
&\quad 0.0822x_{\mathrm{DD}} + 0.0187x_{\mathrm{IBM}} = 0.0444\\
&\quad 0.0701x_{\mathrm{DD}} + 0.1874x_{\mathrm{IBM}} = 0.0335
\end{aligned}
$$

or

$$
\begin{aligned}
x_{\mathrm{DD}} &= 0.5401 - 0.2275 x_{\mathrm{IBM}} \\
x_{\mathrm{IBM}} &= 0.1788 - 0.3741 x_{\mathrm{DD}} \\
x_{\mathrm{DD}} &= 0.5401 - 0.2275(0.1788 - 0.3741 x_{\mathrm{DD}}) \\
x_{\mathrm{DD}} &= 0.5401 - 0.0407 + 0.0851 x_{\mathrm{DD}} \\
x_{\mathrm{DD}} &= 0.4994 + 0.0851 x_{\mathrm{DD}} \\
0.9149 x_{\mathrm{DD}} &= 0.4994 \\
x_{\mathrm{DD}} &= 0.5459 \\
x_{\mathrm{IBM}} &= 0.1788 - 0.3741(0.5459) \\
&= 0.1788 - 0.2042 = -0.0254 \\
X_{\mathrm{D}} &= 1 - x_{\mathrm{DD}} - x_{\mathrm{IBM}} = 1 - 0.5459 - (-0.0254) \\
&= 0.4795
\end{aligned}
$$

One shorts IBM. The risk-minimizing three-asset portfolio expected return is 6.51%.

$$
\begin{aligned}
E\left(R_{\mathrm{p}}\right) &= 0.5459\,(0.0055) - 0.0254\,(0.1107) + 0.4795\,(0.1354) \\
&= 0.0030 - 0.0028 + 0.0649 = 0.0651 \\
\sigma_{\mathrm{P}}^2 &= (0.5459)^2\,(0.1241)^2 + (-0.0254)^2\,(0.2809)^2 + (0.4795)^2\,(0.1365)^2 \\
&\quad +2\,(0.5459)\,(-0.0254)\,(0.3147)\,(0.1241)\,(0.2809) \\
&\quad +2\,(0.5459)\,(0.4795)\,(-0.2108)\,(0.1241)\,(0.1365) \\
&\quad +2\,(-0.0254)\,(0.4795)\,(0.05)\,(0.2809)\,(0.1365) \\
&= 0046 + 0.0001 + 0.0043 + (-0.0003) + (-0.0019) + (-0.0000) \\
&= 0.0068 \\
\sigma_{\mathrm{p}} &= \sqrt{0.0068} = 0.0825
\end{aligned}
$$

The risk-minimizing expected variance is 8.25%.

$$
\mathrm{CV}_{\mathrm{p}}^{\mathrm{EQ}} = \frac{0.1196}{0.0838} = 1.427
$$

$$
\mathrm{CV}_{\mathrm{p}} = \frac{0.0825}{0.0651} = 1.267
$$

The coefficient of variation is considerably lower in the optimal portfolio than in the equally weighted portfolio, as one would have expected. The purpose of Markowitz analysis is to produce the lowest possible level of risk for a given level of return.

Monthly data on stock returns can be found on the University of Chicago Center for Research in Security Prides (CRSP) database from 1926. Let us create a portfolio for GE, D, and IBM using five years of monthly CRSP data from 2002 to 2006, where

Stock	E(R)	(σ_j)		ρ_{ij}	
GE	0.0026	0.0561	1.000	0.1182	0.5305
D	0.0103	0.0531	0.1182	1.000	0.1787
IBM	0.0004	0.0825	0.5305	0.1787	1.000

$$
\begin{aligned}
\frac{\partial \sigma_{\mathrm{p}}^2}{\partial x_1} &= 2x_1\left(\sigma_1^2+\sigma_3^2-2\sigma_{13}\right)+x_2\left(2\sigma_3^2+2\sigma_{12}-2\sigma_{13}-2\sigma_{23}\right)-2\sigma_3^2+2\sigma_{13}=0\\
&= 2x_1\left[(0.0561)^2+(0.0825)^2-2\,(0.0561)\,(0.0825)\,(0.5305)\right]\\
&\quad + x_2\Big[2\,(0.0825)^2+2\,(0.0561)\,(0.0531)\,(0.1182)\\
&\quad -2\,(0.0561)\,(0.0825)\,(0.5305)\ -2\,(0.0531)\,(0.0825)\,(0.1787)\Big]\\
&\quad -2\,(0.0825)^2+2\,(0.0561)\,(0.0825)\,(0.5305)-0.0136+0.0049=0\\
&= 2x_1\,[0.0031+0.0068-0.0049]+x_2\,[0.0136+0.0007-0.0049-0.0016]\\
&= 2x_1\,(0.0050)+x_2\,(0.0078)-0.0087=0\\
&= x_1\,(0.010)+0.0078x_2-0.0087=0
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \sigma_{\mathrm{p}}^2}{\partial x_2} &= 2x_2(\sigma_2^2+\sigma_3^2-2\sigma_{23})+x_1(\sigma_3^2+2\sigma_{12}-2\sigma_{13}-2\sigma_{23})-2\sigma_3^2+2\sigma_{23}=0\\
&= 2x_2\left[(0.0531)^2+(0.0825)^2-2\,(0.0531)\,(0.0825)\,(0.1787)\right]\\
&\quad + x_1\Big[2\,(0.0825)^2+2\,(0.0561)\,(0.0531)\,(0.1182)-2\,(0.0561)\,(0.0825)\\
&\qquad (0.5305)-\ 2\,(0.0531)\,(0.0825)\,(0.1787)\Big]\\
&\quad -2\,(0.0825)^2+2\,(0.0531)\,(0.0825)\,(0.1787)=0\\
&= 2x_2\,(0.0028)+0.0068-0.0016)+x_1\,[0.0136+0.0007-0.0049-0.0016]\\
&\quad -0.0136+0.0016=2x_2\,(0.0080)+x_1\,(0.0078)-0.0120=0\\
&\quad 0.016x_2+0.0078x_1-0.0120=0\\
&\quad 0.016x_2+0.0078x_1=0.012
\end{aligned}
$$

We now have

$$
\begin{aligned}
&0.010x_1+0.0078x_2=0.0087\\
&0.016x_2+0.0078x_1=0.0120\\
&x_1=-\frac{0.0078}{0.0100}x_2+\frac{0.0087}{0.0100}\\
&x_1=-0.78x_2+0.8700
\end{aligned}
$$

$$0.016x_2 + 0.0078\,(-0.78x_2 + 0.8700) = 0.0120$$
$$0.016x_2 - 0.0061x_2 + 0.0068 = 0.0120$$
$$0.0099x_2 = 0.0052$$
$$x_2 = \frac{0.0052}{0.0099} = 0.5263$$
$$x_1 = -0.78\,(0.5253) + 0.87 = -0.4097 + 0.873 = 0.4603$$
$$x_3 = 1 - x_1 - x_2 = 1 - 0.5253 - 0.4603 = 0.0144$$

The equally weighted CRSP GE, D, and IBM portfolio has an annualized expected return of:

$$E(R_\mathrm{p}) = 0.333\,(0.0026) + 0.333\,(0.0103) + 0.333\,(0.0004)$$
$$= 0.0009 + 0.0034 + 0.0001 = 0.0044$$

(5.28% annualized)

and the annualized standard deviation of the portfolio is:

$$\sigma_\mathrm{p}^2 = (0.333)^2\,(0.0561)^2 + (0.333)^2\,(0.0531)^2 + (0.333)^2\,(0.0825)^2$$
$$+ 2\,(0.333)\,(0.333)\,(0.0561)\,(0.0531)\,(0.1182) + 2\,(0.333)\,(0.333)\,(0.0561)$$
$$(0.0825)\,(0.5305) + 2\,(0.333)\,(0.333)\,(0.0531)\,(0.0825)\,(0.1787)$$
$$\sigma_\mathrm{p}^2 = 0.0003 + 0.0003 + 0.0008 + 0.0008 + 0.0005 + 0.0002$$
$$= 0.0029$$
$$\sigma_\mathrm{p} = 0.0539$$

(18.67% annualized)

The coefficient of variation of the equally weighted GE, D, and IBM portfolio is 3.536.

The optimally weighted CRSP GE, D, and IBM have an annualized expected return of:

$$E\left(R_\mathrm{p}\right) = 0.5253\,(0.0026) + 0.4603\,(0.0103) + 0.0144\,(0.0004)$$
$$= 0.0014 + 0.0047 + 0.0000 = 0.0061$$

(7.32% annualized)

and an annualized standard deviation of:

$$\sigma_\mathrm{p}^2 = (0.5253)^2\,(0.0561)^2 + (0.4603)^2\,(0.0531)^2 + (0.0144)^2\,(0.0825)^2$$
$$+ 2\,(0.5253)\,(0.4603)\,(0.0561)\,(0.0531)\,(0.1182)$$
$$+ 2\,(0.5253)\,(0.0144)\,(0.0561)\,(0.0825)\,(0.5303)$$
$$+ 2\,(0.4603)\,(0.0144)\,(0.0531)\,(0.0825)\,(0.1787)$$
$$= 0.0009 + 0.0006 + 0.0000 + 0.0002 + 0.0000 + 0.0000$$
$$\sigma_\mathrm{p}^2 = 0.0017$$
$$\sigma_\mathrm{p} = 0.0412$$

(14.27% annualized)

The coefficient of variation is 1.949, far lower than equally weighted GE, IBM, and D portfolio.

One sees more benefit to diversification in the three security CRSP-monthly data example than in the corresponding *S&P Stock Guide* example. The correlation of GE and IBM is 0.5305 in the CRSP data whereas the corresponding correlation in the S&P annual data is 0.8370. Markowitz diversification involves minimizing correlation coefficients and covariances.

Finally, let us further assume a covariance matrix of securities with lower off-diagonal correlations.

Stock	$E(R_i)$	(σ_i)		ρ_{ij}	
F	−0.0034	0.1145	1	−0.0227	0.0931
JNJ	0.0042	0.0405	−0.0227	1	0.0220
COP	0.0186	0.0632	0.0931	0.0220	1

$$
\begin{aligned}
\frac{\partial \sigma_{\mathrm{p}}^2}{\partial x_1} &= 2x_1\left[(0.1145)^2 + (0.0632)^2 - 2\,(0.1145)\,(0.0632)\,(0.0931)\right]\\
&\quad + x_2\Big[2\,(0.0632)^2 + 2\,(0.1145)\,(0.0405)\,(-0.0227)\\
&\quad - 2\,(0.1145)\,(0.0632)\,(0.0931) - 2\,(0.0405)\,(0.0632)\,(0.0220)\Big]\\
&\quad - 2\,(0.0632)^2 + 2\,(0.1145)\,(0.0632)\,(0.0931) = 0\\
&= 2x_1\,[0.0131 + 0.0040 - 0.0013] + x_2\,[0.0080 - 0.0002 - 0.0013 - 0.0001]\\
&\quad -0.0080 + 0.0013 = 0\\
&= (0.0316)\,x_1 + 0.0064x_2 = 0.0067\\
\frac{\partial \sigma_{\mathrm{p}}^2}{\partial x_2} &= 2x_2\left[(0.0405)^2 + (0.0632)^2 - 2\,(0.0405)\,(0.0632)\,(0.0220)\right]\\
&\quad + x_2[2\,(0.0632)^2 + 2\,(0.1145)\,(0.0405)\,(-0.0227)\\
&\quad - 2\,(0.1145)\,(0.0632)\,(0.0931) - 2\,(0.0405)\,(0.0632)\,(0.0220)]\\
&\quad - 2\,(0.0632)^2 + 2\,(0.0405)\,(0.0632)\,(0.0220) = 0\\
&= 2x_2\,(0.0016 + 0.0040 - 0.0001)\; + x_1\,(0.0080 - 0.0002\\
&\quad -0.0013 - 0.0001) - 0.0080 + 0.0001\\
&= x_2\,(0.0110) + x_1\,(0.0064) - 0.0079 = 0\\
&= 0.0110x_2 + 0.0064x_1 = 0.0079
\end{aligned}
$$

We have

$$
\begin{aligned}
0.0316x_1 + 0.0064x_2 &= 0.0067\\
0.0110x_2 + 0.0064x_1 &= 0.0079
\end{aligned}
$$

Solving:

$$
\begin{aligned}
&0.0110x_2 = -0.0064x_1 + 0.0079\\
&x_2 = -\frac{0.0064}{0.0110}x_1 + \frac{0.0079}{0.0110}\\
&x_2 = -0.5818x_1 + 0.7182\\
&0.0316x_1 + 0.0064\,(-0.5818x_1 + 0.7182) = 0.0067\\
&0.0316x_1 - 0.0037x_1 + 0.0046 = 0.0067\\
&0.0279x_1 = 0.0021\\
&x_1 = 0.0753\\
&x_2 = -0.5818\,(0.0753) + 0.7182 = 0.6744\\
&x_3 = 1 - x_1 - x_2 = 1 - 0.0743 - 0.6744 = 0.2503
\end{aligned}
$$

The equally weighted F, JNJ, and COP portfolio is

$$
\begin{aligned}
E(R_\text{p}) &= 0.333\,(-0.0034) + 0.333\,(0.0042) + 0.333\,(0.0186)\\
&= -0.0011 + 0.0014 + 0.0062 = 0.0065
\end{aligned}
$$

$$(0.0065 \times 12 = 7.80\% \text{ annualized})$$

The standard deviation of the portfolio is

$$
\begin{aligned}
\sigma_\text{p}^2 &= (0.333)^2\,(0.1145)^2 + (0.333)^2\,(0.0405)^2 + (0.333)^2\,(0.0632)^2\\
&\quad + 2\,(0.333)\,(0.333)\,(0.1145)\,(0.0405)\,(-0.0227)\\
&\quad + 2\,(0.333)\,(0.333)\,(0.1145)\,(0.0632)\,(0.0931)\\
&\quad + 2\,(0.333)\,(0.333)\,(0.0405)\,(0.0632)\,(0.0220)\\
&= 0.0015 + 0.002 + 0.0004 - 0.0000 + 0.0001 + 0.000\\
&= 0.0040\\
\sigma_\text{p} &= \sqrt{0.0040} = 0.0632
\end{aligned}
$$

The standard deviation is 6.32%. Thus, the equally weighted F, JNJ, and COP portfolio is 7.80% and the standard deviation is 6.32%, producing a coefficient of variation of 0.810.

The optimally weighted F, JNJ, and COP portfolio has an expected return of 8.64%, and a standard deviation of 6.16%.

$$
\begin{aligned}
E(R_\text{p}) &= 0.0753\,(-0.0034) + 0.6744\,(0.0042) + 0.2503\,(0.0186)\\
&= -0.0003 + 0.0028 + 0.0047 = 0.0072\,(8.64\%)\\
\sigma_\text{p}^2 &= (0.0753)^2\,(0.1145)^2 + (0.6744)^2\,(0.0405)^2 + (0.2503)^2\,(0.0632)^2\\
&\quad + 2\,(0.0753)\,(0.6744)\,(0.1145)\,(0.0405)\,(-0.0227)\\
&\quad + 2\,(0.0753)\,(0.2503)\,(0.1145)\,(0.0632)\,(0.0931)\\
&\quad + 2\,(0.6744)\,(0.2503)\,(0.0405)\,(0.0632)\,(0.0220)
\end{aligned}
$$

$$\sigma_{\mathrm{p}}^2 = 0.0001 + 0.0007 + 0.003 - 0.0000 + 0.0000 + 0.0000$$
$$\sigma_{\mathrm{p}}^2 = 0.0038$$
$$\sigma_{\mathrm{p}} = \sqrt{0.0038} = 0.0616.$$

Thus, the optimally weighted F, JNJ, and COP portfolio is 8.64% and the standard deviation is 6.16%, producing a coefficient of variation of 0.713. At the risk of boring the reader with too many examples, one clearly sees the value of calculating optimal weights for Markowitz diversification and portfolio construction.

1.5 Markowitz Optimization Analysis

The basis form of the Markowitz portfolio optimization analysis can be expressed graphically as by Harry in Bloch et al. (1993). What are the decision variables in Markowitz optimization analysis? One can vary the period of volatility calculation, such as using 5 years of monthly data in calculating the covariance matrix, as was done in Bloch et al. (1993), or 1 year of daily returns to calculate a covariance matrix, as was done in Guerard et al. (1993), or 2–5 years of data to calculate factor returns as in the BARRA system, discussed in Chaps. 2 and 3, and discussed in Menchero et al. (2009).

The goal in investment research is to operate as closely to the Efficient Frontier was possible. The Efficient Frontier, shown in Fig. 1.4, depicts the maximum return for a given level of risk. The Efficient Frontier concept and chart is taught in almost every undergraduate and MBA finance course. Brealey, Myers, and Allen (2006) is perhaps the best example. Applied investment books, such as Farrell (1997) feature the Efficient Frontier.

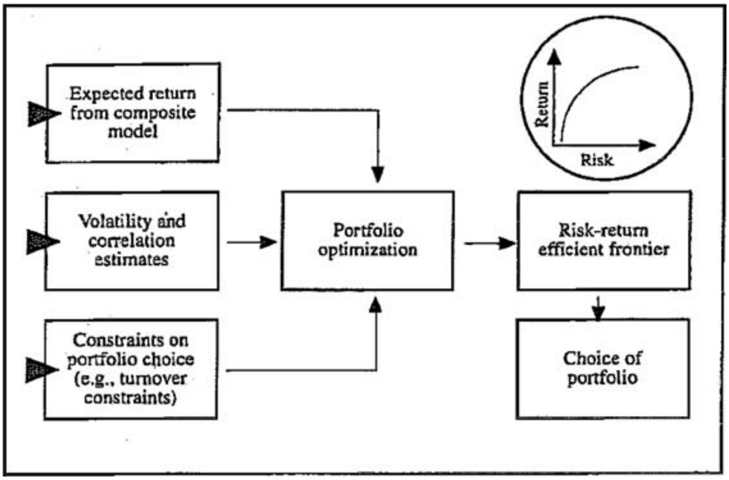


Fig. 1.4 Observed risk-return tradeoff. TSE Section 1, standard composite model. (Source: Bloch et al. 1993)

In Bloch et al. (1993), Markowitz and his research department estimated Efficient Frontiers for Japanese-only securities for the 1974–1990 period. The model is discussed more fully in Chap. 2.

1.5.1 A General Form of Portfolio Optimization

The general four-asset linear programming has been written by Martin (1955) as

$$x_1 + x_2 + x_3 + x_4 = 1 \tag{1.26}$$

$$x_1 \geqslant 0, \mathrm{I} = 1, \ldots, 4$$

$$E = x_1\mu_1 + x_2\mu_2 + x_3\mu_3 + x_4\mu_4 \tag{1.27}$$

E = expected returns

$$\begin{aligned} V = {} & x_1^2\sigma_{11} + x_2^2\sigma_{12} + 2x_1x_2\sigma_{12} + 2x_1x_3\sigma_{13} + 2x_1x_4\sigma_{14} \\ & + x_2^2\sigma_{33} + 2x_3x_4\sigma_{34} + x_4^2\sigma_{44} \end{aligned} \tag{1.28}$$

$$\phi = V + \lambda_1\left(\sum_{i=1}^{4} x_i\mu_i - E\right) + \lambda_2\left(\sum_{i=1}^{4} x_i - 1\right) \tag{1.29}$$

$$\frac{\partial\phi}{\partial x_1} = x_1\sigma_{11} + x_2\sigma_{12} + x_3\sigma_{13} + x_4\sigma_{14} + {}^1\!/\!_2\lambda_1\mu_1 + {}^1\!/\!_2\lambda_2 = 0$$

$$\frac{\partial\phi}{\partial x_1} = x_1\sigma_{14} + x_2\sigma_{24} + x_3\sigma_{34} + x_4\sigma_{44} + {}^1\!/\!_2\lambda_4\mu_4 + {}^1\!/\!_2\lambda_2 = 0$$

The four-asset optimization problem is expressed in a six-equation format by Martin (1955). All partial derivatives are set equal to zero as we did with Markowitz (1959) and the security weights sum to one. The investor is fully invested. The portfolio expected returns is a weighted sum of security expected returns and security weights.

The final set of Markowitz four-asset optimization equations may be written as

$$\begin{aligned} & x_1\sigma_{11} + x_2\sigma_{12} + x_3\sigma_{13} + x_4\sigma_{14} + {}^1\!/\!_2\lambda_1\mu_1 + {}^1\!/\!_2\lambda_2 = 0 \\ & x_1\sigma_{12} + x_2\sigma_{22} + x_3\sigma_{23} + x_4\sigma_{24} + {}^1\!/\!_2\lambda_1\mu_2 + {}^1\!/\!_2\lambda_2 = 0 \\ & \mathrm{x}_1\sigma_{13} + \mathrm{x}_2\sigma_{23} + \mathrm{x}_3\sigma_{33} + \mathrm{x}_4\sigma_{34} + {}^1\!/\!_2\lambda_1\mu_3 + {}^1\!/\!_2\lambda_2 = 0 \\ & \mathrm{x}_1\sigma_{14} + \mathrm{x}_2\sigma_{24} + \mathrm{x}_3\sigma_{34} + \mathrm{x}_4\sigma_{44} + {}^1\!/\!_2\lambda_1\mu_4 + {}^1\!/\!_2\lambda_2 = 0 \\ & \mathrm{x}_1\mu_1 + \mathrm{x}_2\mu_2 + \mathrm{x}_3\mu_3 + \mathrm{x}_4\mu_4 = \mathrm{E} \\ & x_1 + x_2 + x_3 + x_4 = 1 \end{aligned}$$

Markowitz (1959, pp. 282–83) discusses the Rational Mean theory of investing as Latané (1959) as doing at the same time. In the long run, the expected return of the natural logarithm, Ln, of $1 + R$, the portfolio return, is represented by

$$\mathrm{Ln}(1+R) = \mathrm{Ln}(1+E) - \frac{1/2V}{(1+E)^2} \tag{1.30}$$

where E = expt (R) and V = variance (R).

The portfolio that maximizes (1.30) must minimize the variance, V, for a given level of return, and will maximize return in the long run. The maximizing return is the Geometric Mean. Support for the establishment of the Geometric Mean investment criterion is found empirically in Young and Trent (1969) and Levy and Markowitz (1976).

$$\mathrm{GM} = \sqrt[n]{\Pi(1+R)}$$

where $\prod$ stands for the product. Thus, one multiplies one plus the portfolio return for n periods and takes the nth root of the cumulative wealth ratio. The mean variance approximation of (1.30) worked very well in approximating the GM for the 1, 4, 8, 16, and 32 stock portfolios in Young and Trent for the 1953–1960 period using 233 stocks.

Levy and Markowitz found that the mean variance approximation worked very well for 149 investment companies during 1958–1967. Does the investor only need be concerned with mean and variance? What about skewness, the third moment, and Kurtosis, the fourth moment?

Is skewness good for an investment manager? Obviously it is, if skewness is positive, which has been the case historically. This is a huge literature on skewness, but it is sufficient to say that the market rewards positive skewness, see Arditti (1967), Arditti and Levy (1975), and Stone and Guerard (2009). In Arditti (1967), returns are positively associated with variances; the higher the variance, the higher must be the return. However, for skewness, as skewness rises, the market requires lower returns because skewness (positive skewness) is desirable. Across all securities in the S&P Composite Index from 1946 to 1963, Arditti (p. 26) found:

$$\begin{aligned}&\text{Returns} = 0.1044 + 0.4221\,\text{Variance} - 0.1677\,\text{Skewness}\\&(\text{s.e.})(0.0074)(0.0831)(0.0435)\end{aligned}$$

The final form of the Martin (1955) representation of Markowitz can be found in Elton et al. (2007) later shown in this chapter.

Modern-day Markowitz analysis, as illustrated in Elton et al. (2007), continues to use Lagrangian multipliers. One seeks to the ratio of portfolio excess return, defined as the expected return of the portfolio less the risk-free rate, divided by the portfolio standard deviation.

$$\theta = \frac{E(R_{\mathrm{p}}) - R_{\mathrm{F}}}{\sigma_{\mathrm{p}}} \tag{1.31}$$

subject to $\sum_{i=1}^{N} w_i = 1$.

We can rewrite (1.32) as

$$\theta = \frac{\sum_{i=1}^{N} w_i (E(R_i) - R_{\mathrm{F}})}{\left[\sum_{i=1}^{N} w_i^2 \sigma_i^2 + \sum_{\substack{i=1 \\ i \neq j}}^{N} \sum_{j=1}^{N} w_i w_j \sigma_{ij} \right]^{1/2}} \tag{1.32}$$

To maximize the portfolio selection problem in (1.27), one again takes the partial derivatives and sets them equal to 0.

$$\begin{aligned} &\frac{\partial \theta}{\partial w_1} = 0, \quad \frac{\partial \theta}{\partial w_2} = 0, \quad \ldots, \quad \frac{\partial \theta}{\partial w_N} = 0 \\ &\frac{\partial \theta}{\partial w_i} = -\left(\lambda w_1 \sigma_{1i} + \lambda w_2 \sigma_{2i} + \lambda w_3 \sigma_{3i} + \cdots + \lambda x_i \sigma_i^2 + \cdots \lambda w_{N-1} \sigma_{N-1i}\right. \\ &\qquad\quad + E(R_i) - R_{\mathrm{F}} = 0 \end{aligned} \tag{1.33}$$

In (1.33), we may define λw to be z, such that

$$E(R_i) - R_F = z_i \sigma_{1i} + z_2 \sigma_{2i} + \cdots + z_{N-1} \sigma_{N-1} + z_N \sigma_{Ni} \tag{1.34}$$

Thus, Elton, Gruber, Brown, and Goetzmann solve a set of simultaneous equations where

$$\begin{aligned} E(R_i) - R_{\mathrm{F}} &= z_i \sigma_1^2 + z_2 \sigma_{12} + z_3 \sigma_3 + \cdots + z_N \sigma_{1N} \\ E(R_2) - R_{\mathrm{F}} &= z_i \sigma_{12} + z_2 \sigma_2^2 + z_3 \sigma_{32} + \cdots + z_N \sigma_{2N} \\ E(R_3) - R_{\mathrm{F}} &= z_i \sigma_{13} + z_2 \sigma_{23} + z_3 \sigma_3^2 + \cdots + z_N \sigma_{3N} \end{aligned} \tag{1.35}$$

$$w_k = \frac{z_k}{\sum_{i-1}^{N} z_i} \tag{1.36}$$

1.6 Conclusions and Summary

In this chapter, we have introduced the reader to the risk-return trade-off analysis that made Harry Markowitz the well-respected investment researcher that he has been since the early 1950s. Markowitz portfolio selection analysis is standard material in undergraduate and MBA courses. In the next chapter, we introduce the reader to the Capital Asset Pricing Model of Sharpe (1964), Lintner (1965), and Mossin (1966) and show how Markowitz analysis extends to an expanded definition of risk. Markowitz research is implemented daily on Wall Street.

References

Arditti, F.D. 1967. "Risk and the Required Return on Equity." *Journal of Finance* 22, 19–36.

Ardiiti, F.D. and H. Levy. 1975. "Portfolio Efficiency Analysis in Three Moments: The Multiperod Case." *Journal of Finance* 30, 797–809.

Bloch, M., J.B. Guerard Jr, H.M. Markowitz, P. Todd, and G-L. Xu. 1993. "A Comparison of Some Aspects of the U.S. and Japanese Equity Markets." *Japan & the World Economy* 5, 3–26.

Brealey, R.A., S.C. Myers, and F. Allen. 2006. *Principles of Corporate Finance*, 8th edition. New York: McGraw-Hill/Irwin, Chapter 8.

Elton, E.J., M.J. Gruber, S.J. Brown, and W. Goetzmann. 2007. *Modern Portfolio Theory and Investment Analysis*, 8th edition. New York: Wiley.

Fama, E.F. 1976. *Foundations of Finance*. New York: Basic Books.

Fama, E.F. and I.D. MacBeth. 1973. "Risk, Return, and Equilibrium: Empirical Tests." *Journal of Political Economy* 81, 607–636.

Farrell, I.L. Jr. 1997. *Portfolio Management: Theory and Applications*. New York: McGraw-Hill/Irwin.

Graham, B. 1950. *The Intelligent Investor*, 5th edition. New York: Harper & Row, Inc..

Graham, B., D. Dodd, and S. Cottle. 1934, 1962. *Security Analysis: Principles and Technique*. New York: McGraw-Hill Book Company. Editions one and four.

Levy, H. and H.M. Markowitz. 1979. "Approximating Expected Utility by a function of Mean ad Variance." *The American Economic Review* 69, 308–317.

Markowitz, H.M. 1952. "Portfolio Selection." *Journal of Finance* 7, 77–91.

Markowitz, H.M. 1959. *Portfolio Selection: Efficient Diversification of Investment*. Cowles Foundation Monograph No.16. New York: Wiley.

Markowitz, H.M. 1976. "Investment in the Long Run: New Evidence for an Old Rule." *Journal of Finance* 31, 1273–1286.

Markowitz, H.M. 2000. *Mean-Variance Analysis in Portfolio Choice 7 and Capital Markets*. New Hope, PA: Frank J. Fabozzi Associates.

Mossin, J. 1966. "Equilibrium in a Capital Asset Market." *Econometrica* 34, 768–783.

Mossin, J. 1973. *Theory of Financial Markets*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Nerlove, M. 1968. "Factors Affecting Differences Among Rates of Return on Investments in Individual Common Stocks." *Review of Economics and Statistics* 50, 312–331.

Rudd, A. and B. Rosenberg. 1979. "Realistic Portfolio Optimization." In *Portfolio Theory, 25 Years After*. Edited by E. Elton and M. Gruber. Amsterdam: North-Holland.

Rudd, A. and B. Rosenberg. 1980. "The 'Market Model' in Investment Management." *Journal of Finance* 35, 597–607.

Sharpe, W.F. 1963. "A Simplified Model for Portfolio Analysis." *Management Science* 9, 277–293.

Sharpe, W.F. 1971. "A Linear Programming Approximation for the General Portfolio Analysis Problem." *Journal of Financial and Quantitative Analysis 6*.

Stone, B.K. 1970. *Risk, Return, and Equilibrium: A General Single-Period Theory of Asset Selection and Capital Market Equilibrium*. Cambridge, MA: MIT Press.

Stone, B.K. 1973. "A Linear Programming Formulation of the General Portfolio Selection Problem." *Journal of Financial and Quantitative Analysis* 8, 621–636.

Stone, B.K. and J.B. Guerard Jr. 2009. "Methodologies for Isolating and Assessing Potential Portfolio Performance of Stock Return Forecast Models with an Illustration." This volume.

Williams, John Burr. 1938. *The Theory of Investment Value*. Cambridge: Harvard University Press.

Young, W.E. and R.H. Trent. 1969. "Geometric Mean Approximation of Individual Security and Portfolio Performance." *The Journal of Financial and Quantitative Analysis* 4, 179–199.

Chapter 2
Markowitz and the Expanding Definition of Risk: Applications of Multi-factor Risk Models

John B. Guerard, Jr.

In Chap. 1, we introduced the reader to Markowitz mean–variance analysis. Markowitz created a portfolio construction theory in which investors should be compensated with higher returns for bearing higher risk. The Markowitz framework measured risk as the portfolio standard deviation, its measure of dispersion, or total risk. The Sharpe (1964), Lintner (1965), and Mossin (1966) development of the Capital Asset Pricing Model (CAPM) held that investors are compensated for bearing not only total risk, but also rather market risk, or systematic risk, as measured by a stock's beta. Investors are not compensated for bearing stock-specific risk, which can be diversified away in a portfolio context. A stock's beta is the slope of the stock's return regressed against the market's return. Modern capital theory has evolved from one beta, representing market risk, to multi-factor risk models (MFMs) with 4 or more betas.

Investment managers seeking the maximum return for a given level of risk create portfolios using many sets of models, based both on historical and expectation data. In this chapter, we briefly trace the evolution of the estimated models of risk and show how risk models enhance portfolio construction, management, and evaluation. Starting from the path-breaking risk estimations of Bill Sharpe (1963, 1964, 1966), and going on to the MFMs of Cohen and Pogue (1967), Barr Rosenberg (1974, 1982), Farrell (1974, 1998), Blin and Bender (1998), and Stone (1974, 2008), estimations of risk have complemented the risk and return analysis of Markowitz. Several of these MFMs are commercially available and well known by institutional investment advisors.

Let us briefly provide a roadmap of this chapter for the reader. First, the reader is introduced to Capital Market Theory and the relationship between the Capital Market Line and the Security Market Line. Second, we estimate betas and discuss the statistical and financial implications of systematic risk. Third, we discuss the historical developments and approaches to estimation of MFMs. We particularly emphasize the BARRA MFM, as it is the risk model most widely used by asset managers. Fourth, we discuss creating portfolios using risk control models in the USA,

J.B. Guerard, Jr. (✉)
McKinley Capital Management, LLC, Anchorage, AK, USA
e-mail: jguerard@mckinleycapital.com

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_2,
© Springer Science+Business Media, LLC 2010

using analysts' forecast data, and in Japan, using a traditional Markowitz-based risk control model. Fifth, we introduce the reader to an alternative risk model that minimizes the tracking error of portfolios for asset managers. Sixth, we introduce the reader to an alternative optimal portfolio weighting scheme that uses an estimated beta as contrasted with the traditional Markowitz total-risk-minimizing weight illustrated in Chap. 1.

Sharpe (1970) discusses capital market theory in which investors purchase or sell stocks based on beliefs, or predictions, of expected returns, standard deviations of returns, and correlation coefficients of returns. Indeed, all investors have identical expectations of the predictions, known as homogeneous beliefs. Investors seek to maximize return while minimizing risk. Investors may lend and borrow as much as they desire at the pure rate of interest, also known as the risk-free rate of interest. Capital market theory holds that once equilibrium is established, then it is maintained. In equilibrium, there is no pressure to change. Sharpe states that capital market theory asks about relationship between expected returns and risk for (1) portfolios; and (2) securities. The implicit question concerns the appropriate measure of risk for (1) a portfolio and (2) a security. The optimal combination of risky securities is the market portfolio, which is the percentage of market value of the security as compared with the total market value of risky securities. The market portfolio includes only risky assets, and the actual return on the market portfolio is the weighted average of the expected returns of the risky securities. A linear line passes from the risk-free interest rate through the market portfolio on a return-risk graph. An individual preference curve, known as an indifference curve, determines where along this line, known as the Capital Market Line, the investor seeks to invest. A conservative investor might seek to lend money at the risk-free rate and invest the remaining funds in the market portfolio. An aggressive investor might borrow money at the risk-free rate and invest more funds than his or her initial endowment in the market portfolio To increase expected return along the Capital Market Line, the investor must accept more risk. Conversely, to reduce risk, an investor must give up expected return. Sharpe (1970) refers to the slope of the Capital Market Line as the price of risk reduction (in terms of decreased expected return). All efficient portfolios must lie along the capital market line where

$$E(R_{\mathrm{p}}) = R_{\mathrm{F}} + r_{\mathrm{e}}\sigma_{\mathrm{p}} \tag{2.1}$$

where $E(R_{\mathrm{p}})$ = expected portfolio return, r_{e} = price of risk reduction for efficient portfolios, R_{F} = pure (risk-free) rate of interest, and σ_{p} = portfolio standard deviation.

The Capital Market Line, (2.1), summarizes the simple (linear) relationship between expected return and risk of efficient portfolios. However, such a relationship does not hold for inefficient portfolios and individual securities.

Sharpe (1970) presents a very reader-friendly derivation of the Security Market Line (SML). Sharpe assumed that total funds were divided between the market portfolio, M, and security i. The investor is fully invested; hence

$$X_{\mathrm{M}} + x_i = 1$$

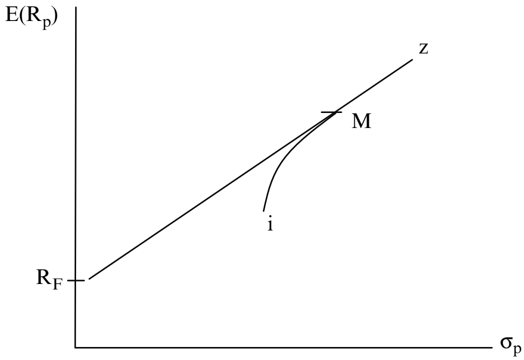


Fig. 2.1 Sharpe's version of the risk-return trade-off analysis

The expected return of the portfolio is

$$E(R_{\rm p}) = x_i E(R_i) + x_{\rm M} E(R_{\rm M})$$

and the corresponding variance of the portfolio, z, is

$$\sigma_{\rm z} = x_i^2 \sigma_i^2 + x_{\rm M}^2 \sigma_{\rm M}^2 + 2 x_i x_{\rm M} \rho_{i{\rm M}} \sigma_i \sigma_{\rm M}$$

We saw these equations in the previous chapter and see no reason for numbering them. Assume that $E(R_i)$ is less than $E(R_{\rm M})$, as is its standard deviation. Sharpe presents this graphically as shown in Fig. 2.1.

We know that the slope of the curve connecting i and M depends upon the relative weights, x_i and $x_{\rm M}$, and the correlation coefficient, $\rho_{i{\rm M}}$, between i and M.

Rewrite $x_{\rm M} = 1 - x_i$ and $\sigma_{i{\rm M}} = \rho_{i{\rm M}} \sigma_i \sigma_{\rm M}$:

$$\sigma_{\rm z} = \sqrt{x_i^2 \sigma_i^2 + (1 - x_i)^2 \sigma_{\rm M}^2 + 2 x_i (1 - x_i) \sigma_{i{\rm M}}}$$

To find a optimal weight:

$$\frac{\partial \sigma_{\rm z}}{\partial x_i} = \frac{x_i \left(\sigma_i^2 + \sigma_{\rm M}^2 - 2\sigma_{i{\rm M}} \right) + \sigma_{i{\rm M}} - \sigma_{\rm M}^2}{\sigma_{\rm z}} \tag{2.2}$$

To minimize portfolio risk, one takes the derivative of total risk relative to the portfolio decision variable, the security weight, as we illustrated in Chap. 1.

$$\begin{aligned} E(R_{\rm z}) &= x_i E(R_i) + (1 - x_i) E(R_{\rm M}) \\ \frac{\partial E_{\rm z}}{\partial x_1} &= E(R_i) - E(R_{\rm M}) \end{aligned} \tag{2.3}$$

One uses (2.2) and the chain rule to give:

$$\frac{\partial E_{\rm z}}{\partial \sigma_{\rm z}} = \frac{\partial E_{\rm z}/\partial x_i}{\partial \sigma_{\rm z}/\partial x_i} = \frac{E(R_i) - E(R_{\rm M})}{\dfrac{x_i \left(\sigma_i^2 + \sigma_{\rm M}^2 - 2\sigma_{i{\rm M}} \right) + \sigma_{i{\rm M}} - \sigma_{\rm M}^2}{\sigma_{\rm z}}} \tag{2.4}$$

Equation (2.4) shows the trade-off between expected returns and portfolio standard deviations. At point M, the market portfolio, $x_i = 0$ and $\sigma_z = \sigma_M$, thus:

$$\left.\frac{\partial E_z}{\partial z_z}\right|_{x_i=0} = \frac{E(R_i) - E(R_M)}{(\sigma_{iM} - \sigma_M^2)/\sigma_M} = \frac{[E(R_i) - E(R_M)]\sigma_M}{\sigma_{iM} - \sigma_M^2} \tag{2.5}$$

In equilibrium, curve iM becomes tangent to the capital Market Line. The investor must be compensated for bearing risk.

The curve iM and slope of the Capital Market Line must be equal. The trade-off of expected return and risk for a security i must be equal to the capital market trade-off.

$$R_e = \sigma_M$$

Thus

$$\frac{[E(R_i) - E(R_M)]\sigma_M}{\sigma_{iM} - \sigma_M^2} = \frac{E(R_M) - R_F}{\sigma_M} \tag{2.6}$$

or

$$E(R_i) - R_F = \left[\frac{E(R_M) - R_F}{\sigma_M^2}\right]\sigma_{iM} \tag{2.7}$$

and

$$E(R_i) = R_F + [E(R_M) - R_F]\frac{\sigma_{iM}}{\sigma_M^2} \tag{2.8}$$

Sharpe (1970) discusses the stock beta as the slope of the firm's characteristic line, the volatility of the security's return relative to changes in the market return.

Beta

The CAPM holds that the return to a security is a function of the security's beta.

$$R_{jt} = R_F + \beta_j[E(R_{Mt}) - R_F] + e_{jt} \tag{2.9}$$

where R_{jt} = expected security return at time t; $E(R_{Mt})$ = expected return on the market at time t; R_F = risk-free rate; β_j = security beta, a random regression coefficient; and e_j = randomly distributed error term.

Let us examine the Capital Asset Pricing Model beta, its measure of systematic risk, from the Capital Market Line equilibrium condition, in an alternative formulation.

$$\beta_j = \frac{\text{Cov}(R_j, R_M)}{\text{Var}(R_M)} \tag{2.10}$$

$$\begin{aligned} E(R_j) &= R_F\left[\frac{E(R_M) - R_F}{\sigma_M{}^2}\right]\text{Cov}(R_j, R_M) \\ &= R_F + [E\ (R_M) - R_F] \end{aligned}$$

$$E(R_j) = R_F + [E(R_M) - R_F]\beta_j \tag{2.11}$$

Table 2.1 IBM Beta

		The REG procedure Model: MODEL1 Dependent variable: ret Analysis of variance			
Source	DF	Sum of squares	Mean square	F value	Pr > F
Model	1	0.2263	0.2263	74.97	<0.0001
Error	58	0.1751	0.0030		
Corrected total	59	0.4013			
	Root MSE	0.0549	R-Square	0.5638	
	Dependent mean	0.0004	Adj R-Sq	0.5563	
	Coeff var	14288			
Parameter estimates					
Variable	DF	Parameter estimate	Standard error	t Value	Pr > $\|t\|$
Intercept	1	−0.0068	0.0071	−0.96	0.3422
sprtrn	1	1.7354	0.2004	8.66	<0.0001

Equation (2.11) defines the Security Market Line (SML), which describes the linear relationship between the security's return and its systematic risk, as measured by beta.

Let us estimate beta coefficients to be used in the Capital Asset Pricing Model (CAPM) to determine the rate of return on equity. One can fit a regression line of monthly holding period returns (HPRs) against the excess returns of an index such as the value-weighted Center for Research in Security Prices (CRSP) index, which is an index of all publicly traded securities. Most security betas are estimated using 5 years of monthly data, some 60 observations, although one can use almost any number of observations. One generally needs at least 30 observations for normality of residuals to occur. One can use the Standard & Poor's 500 Index, the Dow Jones Industrial Index (DJIA), or many other stock indexes. As Table 2.1 shows, IBM's beta is 0.95 when estimated using the equal-weighted CRSP index, 1.64 using the CRSP value-weighted index, and 1.74 using the S&P 500 index, when estimated over the January 2002 to December 2006 period.

The reader sees that the IBM value-weighted CRSP beta has a t-value of 8.66, which is highly statistically significant. One must be careful, because the t-value allows one to reject a null hypothesis that the beta is zero. If the reader wants to calculate a t-value to test the hypothesis that the beta is equal to one, a more plausible assumption, then a second t-value estimated must be calculated. One must divide the beta coefficient, less its hypothesized value, by the estimated standard error of the coefficient.

$$t\text{-value} = \frac{1.74 - 1.00}{0.200} = 3.70$$

The calculated t-value of IBM of 3.70 allows the reader to reject the null hypothesis that the IBM beta is 1.00 at the 5% level (the critical t-value is, of course, 1.96). Thus, IBM's beta is statistically significantly different from 1.0. The stock

Table 2.2 Five-year monthly betas, 1/2002–12/2006

Ticker	S&P500 Beta	t(Beta)	CRSP VW Beta	t(Beta)	CRSP EW Beta	t(Beta)
IBM	1.74	8.66	1.64	7.85	0.95	4.88
GE	0.81	4.54	0.71	3.92	0.31	2.03
D	0.53	2.93	0.54	2.99	0.41	2.91
DD	1.03	6.23	0.94	5.49	0.44	3.30
BP	0.59	3.41	0.61	3.66	0.46	3.46
F	1.84	5.33	1.88	5.61	1.51	5.96
PFE	0.63	3.28	0.58	3.03	0.34	2.20
HWP	1.75	6.90	1.72	6.83	1.22	5.84
RD	0.82	4.34	0.81	4.36	0.51	3.36
JNJ	0.32	2.22	0.25	1.72	0.01	0.09

Table 2.3 Three-year monthly betas, 1/2004–12/2006

Ticker	S&P500 Beta	t(Beta)	CRSP VW Beta	t(Beta)	CRSP EW Beta	T(Beta)
IBM	1.82	5.20	1.54	4.96	0.93	4.35
GE	0.54	1.88	0.37	1.46	0.19	1.16
D	0.12	0.33	0.10	0.31	0.04	0.18
DD	1.23	3.27	0.89	2.59	0.36	1.52
BP	0.40	1.01	0.51	1.51	0.38	1.73
F	1.81	2.71	1.67	2.95	1.05	2.80
PFE	0.74	1.56	0.74	1.56	0.25	0.92
HWP	1.60	3.49	1.56	4.12	1.01	4.02
RD	0.46	1.00	0.49	1.26	0.21	0.83
JNJ	0.04	0.14	−0.10	−0.46	−1.29	0.09

betas reported in the S&P *Stock Guide* use 5 years of monthly returns and the Standard & Poor's 500 Index as their benchmark is 1.72. IBM is an aggressive security, having a beta exceeding unity. A defensive security has a beta less than one. If the (S&P 500) market is expected to rise 10% in the coming year, we should expect IBM stock to rise about 17.3% The corresponding betas are given in Tables 2.2 and 2.3.

If an investor estimates that the security's expected return exceeds the required rate of return from the CAPM and its beta, then the security is considered attractive and it should be purchased. The incremental purchase of the security drives up its price, and lowers its expected return. In this way, equilibrium is established and maintained.

In Chap. 1, we discussed the estimation of an Efficient Frontier in which we maximized expected return relative to risk. An alternative version of an efficient frontier was put forth by Bill Sharpe (1970) and involved estimated betas and expected returns. Here, the useful objective may be written as

$$\text{maximize} \quad Z = (1-\lambda)E_{\text{p}} - \lambda b_{\text{p}} \tag{2.12}$$

Consider the case in which λ equals zero: Z equals E_{p}, and maximizing Z is equivalent to maximizing E_{p}. Combinations of E_{p} and b_{p} lying along any given line have the same value of Z. Higher lines represent larger values of Z. When λ equals zero, the portfolio with the largest attainable value of Z is the one lying at the upper right-hand end of the efficient border. In the King's English, a lambda value of zero indicates that one maximizes expected returns without risk considerations. The portfolio with the largest attainable value of Z plots at the lower left-hand end of the efficient border. If lambda equals one, then one minimizes risk, as measured by the estimated beta. The meaning of λ should be clear: it indicates the relative importance of risk vis-à-vis expected return. To find the complete set of efficient portfolios, one maximizes Z for all values of λ between 0 and 1 (inclusive).

Since all constraints are linear and the objective function is also linear (due to the selection of b_{p} as a measure of risk), the efficient border is simply a series of linear segments. This means that several values of λ may lead to the choice of the same portfolio, while other values may lead to the choice of several alternative portfolios.

Substituting the definitions of E_{p} and b_{p} into the objective function,

$$\begin{aligned} Z &= (1-\lambda)(X_1E_1 + X_2E_2 + \cdots + X_NE_N) \\ &= -(X_1b_1 + X_2b_2 + \cdots + X_NE_N) \end{aligned}$$

Regrouping,

$$Z = X_1[1-\lambda)E_1 - \lambda b_1] + X_2[1-\lambda)E_2 - \lambda b_2] + \cdots + X_N[(1-\lambda)E_N - \lambda b_N]$$

To simplify the notation, let

$$z_i = (1-\lambda)E_i - \lambda b_i$$

Then

$$Z = X_1z_1 + X_2z_2 + \cdots + X_Nz_N \tag{2.13}$$

For any given value of λ, a measure of the investor's tolerance of risk, Sharpe's figure of merit (z_i) can be computed for each security. The higher z_i score, the more desirable the security. If an investor or manager imposed no upper bounds on holdings, then the optimal portfolio would simply consist of the security with the largest value of z_i. Given upper bounds, the solution involves the maximum allowable holding of the most attractive security, plus the maximum allowable holding of the next most attractive until all funds have been invested.

The simplest case arises when the proportion invested in each security can be no more than $1/n$, where n is an integer. For example, assume that n equals 50. Then

$$\begin{aligned} &X_1 \le 0.02 \\ &X_2 \le 0.02 \\ &\vdots \\ &X_N \le 0.02 \end{aligned}$$

For any given value of λ, the solution to such a problem would be to hold the 50 most attractive securities, each one at its upper bound.

In general, to maximize Z for any given value of λ when holdings are constrained so that:

$$X_i \leq \frac{1}{n} \text{ for every security } i \text{ (where } n \text{ is an integer)}$$

1. Compute $z_i = (1 - \lambda)E_i - \lambda b_i$ for every security.
2. Select the n securities with the largest values of z_i.

Each security is represented by a line relating the value of z_i to that of λ. When $\lambda = 0$, $z_i = E_i$, the vertical axis on the left can thus be regarded as plotting the expected return of each security, which is to be maximized. When $\lambda = 1$, $z_i = -b_i$, and risk is minimized. The vertical axis on the right can thus be regarded as plotting the negative of the responsiveness of each security. To draw the line associated with a security, plot and connect the points representing its expected return and responsiveness. In general, values of λ for which the border line is unique give portfolios lying at corners along the efficient E_p, b_p boundary. All other efficient portfolios are simply combinations of adjacent corner portfolios. The graphical representation of the Sharpe Z function is often referred to as the "Sharpe Responsiveness Model."

The difficulty of measuring beta and its corresponding Security Market Line gave rise to extra-market measures of risk, found in the work of King (1966), Farrell (1973), Rosenberg (1973, 1976, 1979), Stone (1974, 2002), Ross (1976) and Ross and Roll (1980), Blin and Bender (1995), and Blin et al. (1998). The BARRA risk model was developed in the series of studies by Rosenberg and completely discussed in Rudd and Clasing (1982) and Grinhold and Kahn (1999) The extra-market risk measures are a seemingly endless source of discussion, debate, and often frustration among investment managers.

In passive portfolio management, deviations of portfolio returns from index returns, often referred to as "tracking error," are expected to be very small. Index funds, in which portfolio weights replicate index (or target) weights, are expected to have very small tracking errors, perhaps 0.1% or lower. Larger tracking errors, possibly reflecting extra-market risks such as returns to illiquidity or size, could impair the reputation of a passive investment manager.

An active portfolio manager, as depicted as a "cowboy" might seek to produce higher portfolio returns by subjecting investors to greater risk. An active manager might have a 6–8% tracking error, leaving anxious investors to drink Maalox or gin. In this chapter, we discuss not only the return-to-risk trade-off, but also how it is assessed by portfolio performance measures.

2.1 Multi-Beta Risk Control Models

Empirical tests of the CAPM often resulted in unsatisfactory results. That is, the average estimated market risk premium was too small, relative to the theoretical market risk premium and the average estimated risk-free rate exceeded the known risk-free rate. Thus low-beta stocks appeared to earn more than was expected and high-beta stocks appeared to earn less than was expected (Black et al. 1972). The equity world appeared more risk-neutral than one would have expected during the 1931–1965 period. There could be many issues with estimating betas using ordinary least squares. Roll (1969, 1977) and Sharpe (1971) identified and tested several issues with beta estimations. Bill Sharpe estimated characteristic lines, the line of stock or mutual fund return vs. the market return, using ordinary least squares (OLS) and the mean absolute deviation (MAD) for the 30 stocks of the Dow Jones Industrial Average stocks vs. the Standard and Poor's 425 Index (S&P 425) for the 1965–1970 period and 30 randomly selected mutual funds over the 1964–1970 period vs. the S&P 425. Sharpe found little difference in the OLS and MAD betas, and concluded that the MAD estimation gains may be "relatively modest."

Perhaps more than one beta should be estimated. Farrell (1974, 1997) estimated a four "factor" model extra-market covariance model. Farrell took an initial universe of 100 stocks in 1974 (due to computer limitations), and ran market models to estimate betas and residuals from the market model:

$$R_{j_t} = a_j + b_j R_{\mathrm{M}t} + e_j \tag{2.14}$$

$$e_{j_t} = R_{j_t} - \hat{a}_j - \hat{b}_j R_{\mathrm{M}_T} \tag{2.15}$$

The residuals of (2.15) should be independent variables. That is, after removing the market impact by estimating a beta, the residual of IBM should be independent of Dow, Merck, or Dominion Resources. The residuals should be independent, of course, in theory. Farrell (1974) examined the correlations among the security residuals of (2.9) and found that the residuals of IBM and Merck were highly correlated, but the residuals of IBM and D (then Virginia Electric & Power) were not correlated. Farrell used a statistical technique known as Cluster Analysis to create clusters, or groups, of securities, having highly correlated market model residuals. Farrell found four clusters of securities based on his extra-market covariance. The clusters contained securities with highly correlated residuals that were uncorrelated with residuals of securities in the other clusters. Farrell referred to his clusters as "Growth Stocks" (electronics, office equipment, drug, hospital supply firms and firms with above-average earnings growth), "Cyclical Stocks" (Metals, machinery, building supplies, general industrial firms, and other companies with above-average exposure to the business cycle), "Stable Stocks" (banks, utilities, retailers, and firms with below-average exposure to the business cycle), and "Energy Stocks" (coal, crude oil, and domestic and international oil firms).

Bernell Stone (1974) developed a two-factor index model which modeled equity returns as a function of an equity index and long-term debt returns. Both equity and debt returns had significant betas. In recent years, Stone et al. (2002, 2008) have developed a portfolio algorithm to generate portfolios that have similar stock betas (systematic risk), market capitalizations, dividend yield, and sales growth cross-sections, such that one can access the excess returns of the analysts' forecasts, forecast revisions, and breadth model, as one moves from low (least preferred) to high (most preferred) securities with regard to your portfolio construction variable (i.e., CTEF or a composite model of value and analysts' forecasting factors). In the Stone et al. (2008) work, the ranking on forecasted return and grouping into fractile portfolios produces a set of portfolios ordered on the basis of predicted return score. This return cross section will almost certainly have a wide range of forecasted return values. However, each portfolio in the cross section will almost never have the same average values of the control variables. To produce a cross-sectional match on any of the control variables, we must reassign stocks. For instance, if we were trying to make each portfolio in the cross section have the same average beta value, we could move a stock with an above-average beta value into a portfolio whose average beta value is below the population average. At the same time, we could shift a stock with a below-average beta value into the above average portfolio from the below-average portfolio. The reassignment problem can be formulated as a mathematical assignment program (MAP).Using the MAP produces a cross-sectional match on beta or any other control variable. All (fractile) portfolios should have explanatory controls equal to their population average value.

Given a cross section of rank-ordered portfolios, the objective of the assignment program is not just to match each portfolio in the cross section on the portfolio average value of beta but to find that particular match that preserves as much as possible a wide range of well-ordered return forecasts while preserving within-portfolio homogeneity of forecasted return subject to the constraints that:

- The portfolio average value of each control variable equal the population mean
- The initial size (number of securities) of each portfolio be preserved
- Each security be fully assigned
- There are no short sales

The crucial constraints are the control matching restrictions. Preserving initial portfolio size and full use of each security are technical constraints that go with full use of the sample. Prohibiting short sales prevents one return observation from canceling out other return observations.

P = number of rank-based portfolios in the cross section
p = 1 is the portfolio with the smallest value of the rank ordering variable
p = P is the portfolio with the largest value of the rank ordering variable
S = total number of securities being assigned to portfolios
s = security subscript
X_{ps} = the fraction of security s assigned to portfolio p
V = representative control variable
VTARGET = target average value of a representative control variable[1]

The advance start that is input to the mathematical assignment program (MAP) is a set of P rank-based fractile portfolios. The linear objective function is a trade-off between preserving as much as possible a measure of cross-portfolio range while minimizing the shifting of stocks away from the original rank-order portfolio. Preserving range and minimizing cross-portfolio mixing are two aspects of statistical

[1] In this study, the target average value is always the ex ante sample average value.

power. They are complementary measures in that optimizing one tends to optimize the other. To reflect the relative importance of these two measures, we define a trade-off parameter Φ that defines a relative weighting, where $0 < \Phi < 1$, reflected by a weighting parameter Φ. Thus, the objective function can be written:

$$\text{Maximize : OBJECTIVE FUNCTION} = \Phi[\text{RANGE}] - (1 - \Phi)[\text{SHIFTING}]$$

Let D_{ps} be the squared difference in the numerical rank of between portfolio p and the natural portfolio rank of security s in the initial range-based partitioning. The set of D_{ps} can be scaled such that large shifts are much worse than small ones. If FS_{s} denotes the value of the forecast score for stock s, then the objective function above can be written in terms of assignment variables as

$$\text{Maximize } \Phi\left[\Sigma_{\text{S}} X_{\text{ps}} FS_{\text{s}} - \Sigma_5 X_{1\text{s}} FS_{\text{s}}\right] - (1 - \Phi)\left[\Sigma_{\text{p}} \Sigma_{\text{s}} X_{\text{ps}} D_{\text{ps}}\right] \tag{2.16}$$

The MAP can be solved for a range of trade-off values by varying Φ from zero to 1. However, experience shows that the solutions are robust to variation in the trade-off Φ. Because these two attributes of statistical power are complementary objectives, minimizing cross-fractile shifting generally preserves the range in the starting fractile portfolios. Note that the Stone et al. analysis represents a multi-surface programming problem in the spirit of the Sharpe Responsiveness Model. This result should not be surprising since Bill Sharpe and Bernell Stone developed linear programming models of portfolio construction as far back as 1973.

Let V_s denote the security s value of a representative control variable. Let VTARGET denote the target value of this representative control variable for all P portfolios in the cross section. The representative control constraint can be expressed as:

$$\sum_{\text{s}} X_{\text{ps}} V_j = V\text{TARGET}_j \quad \text{p} = 1, \ldots, \text{P} - 1 \tag{2.17}$$

We impose two generic data usage constraints. The first says that each security must be fully assigned to one or more portfolios, i.e.,

$$\sum_{\text{p}} X_{\text{ps}} = 1, \quad \text{s} = 1, \ldots, \text{S} \tag{2.18}$$

The second security assignment constraint keeps the number of securities in each matched portfolio the same as the number of securities in the corresponding fractile of the starting rank-order partitioning of the distribution of V1. Let F_{p} denote the number of securities in fractile p. Then this restriction is

$$\sum_{\text{s}} X_{\text{ps}} = F_{\text{p}}, \quad \text{p} = 1, \ldots, \text{P} \tag{2.19}$$

The no short-sale restriction and the natural limitation that no security can be used more than once requires:

$$0 \leqslant X_{\text{ps}} \leqslant 1 \tag{2.20}$$

The substance of the reassignment process is well understood by knowing input and output. The input is a cross section formed by ranking stocks into fractile portfolios, which is the focus of most cross-sectional return analyses in past work on cross-sectional return dependencies. The output is a cross section of fractile portfolios that are matched on a specified set of explanatory controls. Optimization arises in finding the particular reassignment that maximizes a trade-off between preserving the widest possible range of well-ordered portfolio values of forecasted return and also ensuring preservation of within-portfolio homogeneity of forecasted return. The original input cross section is transformed by the optimal reassignment of stocks to produce a new cross section that is matched on the average values of a set of explanatory controls.[2] Stone and Guerard (2009) tested the hypothesis that a composite model of earnings, book value, cash flow, and sales-based value-investing strategies produces superior returns even after correcting for risk, growth, and other return impact variables. The Stone and Guerard (2009) paper is included in this volume.

In 1976, Ross published his "Arbitrage Theory of Capital Asset Pricing," which held that security returns were a function of several (4–5) economic factors. Ross and Roll (1980) empirically substantiated the need for 4–5 factors to describe the return generating process. In 1986, Chen, Ross, and Roll developed an estimated multi-factor security return model based on

$$R = a + b_{\mathrm{MP}}\mathrm{MP} + b_{\mathrm{DEI}}\mathrm{DEI} + b_{\mathrm{UI}}\mathrm{UI} + b_{\mathrm{UPR}}\mathrm{UPR} + b_{\mathrm{UTS}}\mathrm{UTS}\ \mathrm{te}_t \tag{2.21}$$

where MP = monthly growth rate of industrial production, DEI = change in expected inflation, UI = unexpected inflation, UPR = risk premium, and UTS = term structure of interest rates.

Chen, Ross, and Roll (CRR) defined unexpected inflation as the monthly (first) differences of the Consumer Price Index (CPI) less the expected inflation rate. The risk premia variable is the "Baa and under" bond return at time and less the long-term government bond return. The term structure variable is the long-term government bond return less the Treasury bill rates, known at time $t-1$, and applied to time t. When CRR applied their five-factor model in conjunction with the value-weighted index betas, during the 1958–1984 period, the index betas are not statistically significant whereas the economic variables are statistically significant. The Stone, Farrell, Dhrymes et al. (1984) and Dhrymes et al. (1985) and Chen, Ross, and Roll multi-factor model used 4–5 factors to describe equity security risk. The models used different statistical approaches and economic models to control for risk. Asset managers create portfolios for clients, investors, who seek to maximize expected portfolio returns relative to risk. The purpose of this section has been to introduce the reader to Capital Market Theory, beta estimations, and the relevance of multi-factor risk control models.

[2] Given our sample data, a decision about the number of portfolios to be formed, and an initial rank-ordering into fractile portfolios, we are maximizing the statistical power to test for return dependencies for portfolio observations that belong to a well-defined subsurface of an overall multivariate cross-sectional return dependency.

2.2 The Barra Model: The Primary Institutional Risk Model

Barr Rosenberg and Walt McKibben (1973) estimated the determinants of security betas and standard deviations. This estimation formed the basis of the Rosenberg extra-market component study (1974), in which security specific risk could be modeled as a function of financial descriptors, or known financial characteristics of the firm. Rosenberg and McKibben found that the financial characteristics that were statistically associated with beta during the 1954–1970 period were:

(1) Latest annual proportional change in earnings per share
(2) Liquidity, as measured by the quick ratio
(3) Leverage, as measured by the senior debt-to-total assets ratio
(4) Growth, as measured by the growth in earnings per share
(5) Book-to-Price ratio
(6) Historic beta
(7) Logarithm of stock price
(8) Standard deviation of earnings per share growth
(9) Gross plant per dollar of total assets
(10) Share turnover

Rosenberg and McKibben used 32 variables and a 578-firm sample to estimate the determinants of betas and standard deviations. For betas, Rosenberg and McKibben found that the positive and statistically significant determinants of beta were the standard deviation of EPS growth, share turnover, the price-to-book multiple, and the historic beta. Rosenberg et al. (1975), Rosenberg and Marathe (1979), Rudd and Rosenberg (1979, 1980), and Rudd and Clasing (1982) expanded upon the initial Rosenberg MFM framework.

In 1975, Barr Rosenberg and his associates introduced the Barra US Equity Model, often denoted USE1. We spend a great deal of time on the Barra USE1 and USE3 models because 70 of the 100 largest investment mangers use the Barra USE3 Model.[3] The BARRA USE1 Model predicted risk, which required the evaluation of the firm's response to economic events, which were measured by the company's fundamentals. There were six descriptors, or risk indexes, in the Barra model. These descriptors were composite variables primary based on the statistically significant variables in Rosenberg and McKibben (1973). Rudd and Clasing (1982) is an excellent reference for how the Barra equity model is constructed. Barra is a proprietary model; that is, the composite model weights are not disclosed. Thus, there were nine factors in the Index of Market Variability, including the historic beta estimate, historic sigma estimate, share turnover for 3 months, trading volume, the log of the common stock price, and a historical alpha estimate, and cumulative range over 1 year, but without coefficients, one cannot reproduce the model. One can correlate an investment manager's variables with the risk indexes, as we will discuss later

[3] According to MSCI/Barra online advertisements. Morgan Stanley Capital International (MSCI) purchased Barra in 2004. The Barra model is recognized as the MSCI Barra model, which will be referred to as simply "Barra" in this volume.

in the chapter. The Index of Earnings Variability included the variance of earnings, variance of cash flow, and the covariability of earnings and price. The Index of Low Valuation and Unsuccess included the growth in earnings per share, recent earnings change, relative strength (a price momentum variable), the book-to-price ratio, dividend cuts, and the return of equity. The Index of Immaturity and Smallness included the log of total assets, the log of market capitalization, and net plant/common equity. The Index of Growth Orientation included the dividends-to-earnings ratio (the payout ratio), dividend yield, growth in total assets, the earnings-to-price (ep) multiple, and the typical ep ratio over the past 5 years. The Graham and Dodd low P/E investment manager would "load up" on The Index of Growth Orientation, and would offer investors positive asset selection (good stock picking) only if the portfolio weights differed from weights on the "Growth" Index components. The Index of Financial Risk" included leverage at market and book values, debt-to-assets ratio, and cash flow-to-current liabilities ratio.[4]

There were 39 industry variables in the Barra USE1 model. How is the data manipulated and/or normalized to be used in the Barra USE1 model? First, raw data is normalized by subtracting a mean and dividing through by the variable standard deviation; however, the mean subtracted is the market capitalization weighted mean for each descriptor for all securities in the S&P 500. Rudd and Clasing remind us that the capitalization weighted value for S&P 500 stocks is zero. The relevant variable standard deviation is not the universe standard deviation of each variable, but the standard deviation of the variables for companies with market capitalizations exceeding $50 million. A final transformation occurs when the normalized descriptor is scaled such that its value is one standard deviation above the S&P 500 mean. Every month the monthly stock returns in the quarter are regressed as a function of the normalized descriptors. If the firm is typical of the S&P 500 firms, then most of the scaled descriptor values and coefficients should be approximately zero. The monthly residual risk factors are calculated by regressing residual returns (the stock excess return less the predicted beta times the market excess return) vs. the six risk indexes and the industry dummy variables.

The statistically significant determinants of the security systematic risk became the basis of the Barra E1 Model risk indexes. The domestic Barra E3 (USE3, or sometimes denoted US-E3) model, with some 15 years of research and evolution, uses 13 sources of factor, or systematic, exposures. The sources of extra-market factor exposures are volatility, momentum, size, size non-linearity, trading activity, growth, earnings yield, value, earnings variation, leverage, currency sensitivity, dividend yield, and non-estimation universe. The Barra USE3 descriptors are included in the appendix to this chapter.

How does USE3 differ from USE1? There are many changes; of importance to many readers, the USE3 uses analysts' predictions of the current year and 1-year-ahead earnings per share in the earnings yield index which is used in conjunction

[4] See Rudd and Clasing (1982), p. 115, for the USE1 descriptors. We use the original Rudd and Clasing names of risk index factors in this chapter.

with the historic and 12-month trailing earnings-to-price multiples. The analysts' standard deviation of forecasts is a component of the earnings variabilitycomponent. Momentum, book-to-market (denoted as "value"), and dividend yield are now separate risk indexes.

The total excess return for a multiple factor model, referred to as the MFM, in the Rosenberg methodology for security j, at time t, dropping the subscript t for time, may be written:

$$E(R_j) = \sum_{k=1}^{K} \beta_{jk} \tilde{f}_k + \tilde{e}_j \tag{2.22}$$

The non-factor, or asset-specific, return on security j, is the residual risk of the security, after removing the estimated impacts of the K factors. The term, f, is the rate of return on factor k. A single factor model, in which the market return is the only estimated factor, is obviously the basis of the Capital Asset Pricing Model. Accurate characterization of portfolio risk requires an accurate estimate of the covariance matrix of security returns. A relatively simple way to estimate this covariance matrix is to use the history of security returns to compute each variance, covariance, and security beta. The use of beta, the covariance of security and market index returns, is one method of estimating a reasonable cost of equity funds for firms. However, the approximation obtained from simple models may not yield the best possible cost of equity. The simple, single index beta estimation approach suffers from two major drawbacks:

- Estimating a covariance matrix for the Russell 3,000 stocks requires a great deal of data
- It is subject to estimation error. Thus, one might expect a higher correlation between DuPont and Dow than between DuPont and IBM, given that DuPont and Dow are both chemical firms

Taking this further, one can argue that firms with similar characteristics, such as firms in their line of business, should have returns that behave similarly. For example, DuPont and IBM will all have a common component in their returns because they would all be affected by news that affects the stock market, measured by their respective betas. The degree to which each of the three stocks responds to this stock market component depends on the sensitivity of each stock to the stock market component.

Additionally, one would expect DuPont and Dow to respond to news effecting the chemical industry, whereas IBM and Dell would respond to news effecting the Computer industry. The effects of such news may be captured by the average returns of stocks in the chemical industry and the computer industry. One can account for industry effects in the following representation for returns:

$$\begin{aligned} \tilde{r}_{\mathrm{DD}} = \;& E[\tilde{r}_{\mathrm{DD}}] + \beta \cdot [\tilde{r}_{\mathrm{M}} - E[\tilde{r}_{\mathrm{M}}]] \\ & +1 \cdot [\tilde{r}_{\mathrm{CHEMICAL}} - E[\tilde{r}_{\mathrm{CHEMICAL}}]] + 0 \cdot [\tilde{r}_{\mathrm{C}} - E[\tilde{r}_{\mathrm{DD}}]] + \mu_{\mathrm{P}} \end{aligned} \tag{2.23}$$

where $\tilde{r}_{\mathrm{DD}}$ = DD's realized return, $\tilde{r}_{\mathrm{M}}$ = the realized average stock market return, $\tilde{r}_{\mathrm{CHEMICAL}}$ = realized average return to chemical stocks, $\tilde{r}_{\mathrm{C}}$ = the realized average return to computer stocks, $E[\cdot]$ = expectations, β_{DD}= DD's sensitivity to stock market returns, and μ_{DD} = the effect of DD specific news on DD returns.

This equation simply states that DD's realized return consists of an expected component and an unexpected component. The unexpected component depends on any unexpected events that affect stock returns in general $[\tilde{r}_{\text{M}} - E[\tilde{r}_{\text{M}}]]$, any unexpected events that affect the chemical industry $[\tilde{r}_{\text{CHEMICAL}} - E[\tilde{r}_{\text{CHEMICAL}}]]$, and any unexpected events that affect DD alone (μ_{DD}). Thus, the sources of variation in DD's stock returns are variations in stock returns in general, variations in chemical industry returns, and any variations that are specific to DD. Moreover, DD and Dow returns are likely to move together because both are exposed to stock market risk and chemical industry risk. DD, IBM, and D, on the other hand, are likely to move together to a lesser degree because the only common component in their returns is the market return.

Investors look at the variance of their total portfolios to provide a comprehensive assessment of risk. To calculate the variance of a portfolio, one needs to calculate the covariances of all the constituent components. Without the framework of a multiple-factor model, estimating the covariance of each asset with every other asset is computationally burdensome and subject to significant estimation errors. Let us examine the risk structure of the Barra MFM.

$$\begin{aligned} \text{V(I,j)} &= \text{Covariance}[\text{r}(\tilde{\text{i}}), \text{r}(\tilde{\text{j}})] \\ \text{where} \quad \text{V(i,j)} &= \text{asset covariance matrix, and} \\ \text{i,j} &= \text{individual stocks.} \end{aligned}$$

$$\text{V} = \begin{bmatrix} \text{V(1,1)} & \text{V(1,2)} & \cdots & \text{V(1,N)} \\ \text{V(2,1)} & \text{V(2,2)} & \cdots & \text{V(2,N)} \\ \text{V(3,1)} & \text{V(3,2)} & \cdots & \text{V(3,N)} \\ \vdots & \vdots & & \\ \text{V(N,1)} & \text{V(N,2)} & \cdots & \text{V(N,N)} \end{bmatrix}$$

Fig. 2.2 The Barra covariance matrix

The Barra MFM simplifies these calculations dramatically, replacing individual company profiles with categories defined by common characteristics (factors). The specific risk is assumed to be uncorrelated among the assets and only the factor variances and covariances are calculated during model estimation. Let us briefly review how Barr Rosenberg initially estimated the BARRA factor structure.

The multiple-factor risk model significantly reduces the number of calculations. For example, in the U.S. Equity Model (USE3), 65 factors capture the risk characteristics of equities. This reduces the number of covariance and variance calculations; moreover, since there are fewer parameters to determine, they can be estimated with greater precision. The Barra risk management system begins with the MFM equation:

$$\tilde{r}_{\text{i}} = X\tilde{f} + \tilde{u} \tag{2.24}$$

$$\tilde{r} = X\tilde{f} + \tilde{u}$$

where $\tilde{r}$ = vector of excess returns,

X = exposure matrix,

$\tilde{f}$ = vector of factor returns, and

$\tilde{u}$ = vector of specific returns.

$$\begin{bmatrix} \tilde{r}\,(1) \\ \tilde{r}\,(2) \\ \vdots \\ \tilde{r}\,(N) \end{bmatrix} = \begin{bmatrix} X(1,1) & X(1,2) & \cdots & X(1,K) \\ X(2,1) & X(2,2) & \cdots & X(2,K) \\ \vdots & \vdots & & \vdots \\ X(N,1) & X(N,2) & \cdots & X(N,K) \end{bmatrix} \begin{bmatrix} \tilde{f}\,(1) \\ \tilde{f}\,(2) \\ \vdots \\ \tilde{f}\,(K) \end{bmatrix} + \begin{bmatrix} \tilde{u}\,(1) \\ \tilde{u}\,(2) \\ \vdots \\ \tilde{u}\,(N) \end{bmatrix}$$

Fig. 2.3 The Barra risk structure

where $\tilde{r}_{\mathrm{i}}$ = excess return on asset i, X = exposure coefficient on the factor, $\tilde{f}$ = factor return, and $\tilde{u}$ = specific return.

Substituting this relation in the basic equation, we find that:

$$\text{Risk} = \text{Var}(\tilde{r}_j) \tag{2.25}$$

$$= \text{Var}(\mathrm{X}\tilde{f} + \tilde{u}) \tag{2.26}$$

Using the matrix algebra formula for variance, the risk equation becomes

$$\text{Risk} = \mathrm{XFX}^{\mathrm{T}} + \Delta \tag{2.27}$$

where X = exposure matrix of companies upon factors, F = covariance matrix of factors, X^{T} = transpose of X matrix, and Δ = diagonal matrix of specific risk variances.

This is the basic equation that defines the matrix calculations used in risk analysis in the Barra equity models.[5] Investment managers seek to maximize portfolio return for a given level of risk. For many managers, risk is measured by the Barra risk model.

In 2008, MSCI Barra introduced the Barra Global Equity Model, Version 2. Jose Menchero and his colleagues at Barra authored the "Global Equity Risk Modeling" a working paper that was revised and is published in this volume. Menchero et al. (2009) estimated an eight-risk index model in the spirit of the Rosenberg USE3 model.

[5] Markowitz discusses the MFM formulation in his second monograph, *Mean-Variance Analysis in Portfolio Choice and Capital Markets*, New Hope, PA: Frank J. Fabozzi Associates, 2000), Chap. 3, pp. 45–47. The Markowitz (1987, 2000) Mean-Variance Analysis volume requires great patience and thought on the part of the reader, as noted by Bill Sharpe in his foreword to the 2000 edition.

2.3 The Barra Multi-factor Model: An Example of Analysts' Forecasts, Revisions, and Breadth

Let us address the Barra Multifactor Model and its risk indexes using estimated earnings forecasting components of an I/B/E/S-based model of (consensus) forecasted earnings, earnings revisions, and the direction of earnings per share EPS revisions of 1-year-ahead (FY1) and 2-year-ahead (FY2) forecasts. Guerard and Mark (2003) referred to the composite earnings forecasting model as the CTEF model. The CTEF model produced not only higher returns, and returns relative to risk, than its components, but also higher and more statistically significant asset selection than its components in the Russell 3000 universe during the 1990–2001 period. This discussion is useful because it covers the issue of variables in the BARRA risk indexes. See Table 2.4 for Russell 3000 earnings component results in which portfolios of approximately 100 stocks are produced by tilting on the individual and component CTEF variables. The forecast earnings for share for the 1-year-ahead and 2-year-ahead periods, FEP1 and FEP2, offer negative, but statistically insignificant asset selection. The total active returns are positive, and not statistically significant. The asset selection is negative because the FEP variables have positive and statistically significant loadings on the risk indexes; particularly the earnings yield index. The factor loading of the FEP variables on the earnings yield risk index is not unexpected, given that the earnings yield factor index in the USE3 includes the forecast earnings-to-price variable. Thus, there is no multiple factor model benefit to the FEP variables. Note that there is no statistically significant rewards to sector variables in the analysts' forecasts.

The monthly revision variables, the RV variables, offer no statistically significant total active returns, or asset selection abilities, as analysts' revisions are incorporated into the USE3 model, as previously mentioned. The breadth variables, BR, produce statistically significant total active returns and asset selection, despite a statistically significant risk index loading. The breadth variable load on the earnings yield and growth risk indexes. Let us take a closer look at the BR1 factor risk index loading. The BR1 variable leads a portfolio manager to have a positive average active exposure to the earnings yield index, which incorporates the analyst predicted

Table 2.4 Components of the composite earnings forecasting variable, 1990–2001 Russell 3000 Universe

R3000 earnings analysis	Total Active	T-stat	Asset Selection	T-stat	Risk Index	T-stat	Sectors	T-stat
FEP1	2.14	1.61	−1.18	−1.17	4.20	4.42	−0.86	−1.34
FEP2	1.21	0.91	−1.43	−1.35	3.33	3.35	−0.78	−1.15
RV1	0.76	0.69	0.34	0.42	0.92	1.46	−0.34	−0.89
RV2	1.40	1.37	1.09	1.31	0.81	1.42	−0.39	−1.08
BR1	2.59	2.83	1.85	2.43	1.08	2.15	−0.20	−0.51
BR2	2.43	2.36	1.51	1.75	1.09	2.04	−0.01	−0.02
CTEF	2.87	2.81	2.07	2.66	1.19	1.70	−0.26	−0.66

earnings-to-price and historic earnings-to-price measures. The BR1 tilt has a negative and statistically significant average exposure to size, non-linearity, defined as the cube of normalized market capitalization. This result is consistent with analyst revisions being more effective in smaller capitalized securities. The BR1 variable tilt leads the portfolio manager to have a positive and statistically significant exposure to the growth factor index, composed of the growth in the dividend payout ratio, the growth rates in total assets and earnings per share during the past 5 years, recent 1 year earnings growth, and the variability in capital structure. The CTEF variable produces statistically significant total active returns and asset selection. The CTEF variable loading on the risk index is statistically significant at the 10% level because of its loading on the earnings yield and non-linear size indexes, as was the case with its breadth components (see Table 2.5). The CTEF model offers statistically significant asset selection in a multiple factor model framework.

The Frank Russell large market capitalization universe (the Russell 1000), middle market capitalization (mid cap), small capitalization (Russell 2000), and small- and middle market capitalization (Russell 2500) universes were used in the Guerard and Mark (2003) CTEF tests. Higher excess returns, greater asset selection, and stock market inefficiency were found in the smaller stocks of the R3000 universe. See Guerard and Mark (2003) for a more complete discussion of the earnings forecasting alpha work within the Barra system. An outstanding survey of the earnings forecasting literature and its use in hypothesis testing in accounting and finance can be found in Ramnath et al. (2008).

Table 2.5 CTEF variable factor exposures, Russell 1000 Universe
Attribution analysis
Annualized contributions to risk index return

Source of return	Average Active Exposure	Contribution (% Return) Average [1]	Variation [2]	Total [1 + 2]	Total Risk (% Std Dev)	Info Ratio	T-Stat
Volatility	−0.01	0.01	−0.07	−0.06	0.17	−0.32	−1.12
Momentum	0.12	−0.07	0.08	0.01	0.60	0.03	0.11
Size	−0.20	0.36	−0.09	0.27	0.93	0.24	0.83
Size non-linearity	−0.02	0.02	0.03	0.05	0.10	0.44	1.52
Trading activity	0.00	0.00	0.01	0.01	0.11	0.11	0.37
Growth	−0.05	0.05	0.03	0.08	0.14	0.48	1.65
Earnings yield	0.13	0.66	−0.12	0.55	0.40	1.20	4.13
Value	0.06	0.03	0.02	0.06	0.17	0.30	1.03
Earnings variation	0.02	−0.02	0.00	−0.03	0.10	−0.21	−0.73
Leverage	0.06	−0.01	−0.04	−0.04	0.17	−0.23	−0.80
Currency sensitivity	−0.02	0.01	−0.05	−0.04	0.11	−0.32	−1.11
Yield	0.04	0.01	−0.04	−0.04	0.14	−0.24	−0.81
Non-EST universe	0.00	0.00	0.01	0.01	0.04	0.20	0.68
Total				0.82	1.16	0.62	2.13

2.4 Tracking Error and Markowitz Portfolio Construction

The estimation of security weights in a portfolio is the primary calculation of Markowitz's portfolio management. Let us take a different look at the question of security weights. The security weight, as we have discussed, is the proportion of the portfolio market value invested in the individual security:

$$W_s = \frac{M v_1}{M v_p} \tag{2.28}$$

where w_s = portfolio weight insecurity s, MV_s = value of security s within the portfolio, and MV_p = the total market value of portfolio.

The active weight of the security is calculated by subtracting the security weight in the (index) benchmark, b, from the security weight in the portfolio, p.

$$w_{s,p} - w_{s,b} \tag{2.29}$$

That is, if IBM has a 3.0% weight in the portfolio while its weight in the index benchmark is only 2.5%, then IBM has a positive (0.5%) active weight in the portfolio. The portfolio manager has an active "bet" on securities on which he or she has a positive active weight and a negative bet on those securities with negative active weights. Markowitz analysis and its efficient frontier minimized risk for a given level of predicted return. The asset manager creates portfolios to maximize expected the geometric mean and the Sharpe Ratio. An asset manager must be aware of the portfolio tracking error relative to a universe benchmark. Risk can be measured by its volatility which can be defined as the standard deviation in the portfolio return over a forecast horizon, which Blin and Bender in the APT Associates Reference Guide (1995) and Blin et al. (1997) normally define as 1 year, as we have noted.

$$\sigma = \sqrt{E(r_p - E(r_p))^2}$$

Let us once again break down volatility into systematic and specific risk.

$$\sigma_p^2 = \sigma_{\beta p}^2 + \sigma_{\varepsilon p}^2$$

where σ_p = Total Portfolio Volatility, $\sigma_{\beta p}$ = Systematic Portfolio Volatility, and $\sigma_{\varepsilon p}$ = Specific Portfolio Volatility.

Blin and Bender created a multi-factor risk model within their APT risk model based on forecast volatility.

$$\sigma_p = \sqrt{52\left(\sum_{C=1}^{c}\left(\sum_{i=1}^{s} w_i \beta_{i,c}\right)^2 + \sum_{i=1}^{s} w_i^2 \varepsilon_{i,w}^2\right)} \tag{2.30}$$

where σ_{p} = Forecast Volatility of Annual Portfolio Return, C = Number of Statistical Components in the Risk Model, w_i = Portfolio weight in security i, $\beta_{i,\mathrm{c}}$ = The loading (beta) of security i on risk component c, $\varepsilon_{i,\mathrm{w}}$ = Weekly Specific Volatility of Security i.

The Blin and Bender systematic volatility is a forecast of the annual portfolio standard deviation expressed as a function of each security's systematic APT components

$$\sigma\beta_{\mathrm{p}} = \sqrt{52 \sum_{C=1}^{\mathrm{c}} \left(\sum_{i=1}^{\mathrm{s}} w_i \beta_{i,\mathrm{c}} \right)^2} \tag{2.31}$$

Portfolio specific volatility is a forecast of the annualized standard deviation associated with each security's specific return.

$$\sigma_{\varepsilon\mathrm{p}} = \sqrt{52 \sum_{i=1}^{\mathrm{s}} w_i^2 \varepsilon_{i,\mathrm{w}}^2} \tag{2.32}$$

The tracking error, te, is a measure of volatility applied to the active return of funds (or portfolio strategies) that are indexed vs. (against) a benchmark, which is often an index. Portfolio tracking error is defined as the standard deviation of the portfolio return less the benchmark return over 1 year.

$$\sigma_{\mathrm{te}} = \sqrt{E\left(((r_{\mathrm{p}} - r_{\mathrm{b}}) - E(r_{\mathrm{p}} - r_{\mathrm{b}}))^2\right)} \tag{2.33}$$

σ_{te} = annualized tracking error, r_{p} = actual portfolio annual return, r_{b} = actual benchmark annual return.

The APT-reported tracking error is the forecast tracking error for the current portfolio vs. the current benchmark for the coming year.

$$\sigma_{\mathrm{te}} = \sqrt{52 \left(\sum_{C=1}^{\mathrm{c}} \left(\sum_{i=1}^{\mathrm{s}} w_{i,\mathrm{p}} - w_{i,\mathrm{b}} \right) \beta_{i,\mathrm{c}} \right)^2 + \sum_{i=1}^{\mathrm{s}} \left(w_{i,\mathrm{p}} - w_{i,\mathrm{b}} \right)^2 \varepsilon_{i,\mathrm{w}}^2} \tag{2.34}$$

where $w_{i,\mathrm{p}} - w_{i,\mathrm{b}}$ = portfolio active weight and the other terms have been defined in (2.29).

Systematic tracking error of a portfolio is a forecast of the portfolio's active (annual) return as a function of the securities' returns associated with APT risk model components.

$$\sigma\beta_{\mathrm{te}} = \sqrt{52 \sum_{C=1}^{\mathrm{c}} \left(\sum_{i=1}^{\mathrm{s}} \left(w_{i,\mathrm{p}} - w_{i,\mathrm{b}} \right) \beta_{i,\mathrm{c}}^2 \right)} \tag{2.35}$$

Portfolio specific tracking error can be written as a forecast of the annual portfolio active return associated with each security's specific behavior.

$$\sigma_{\varepsilon \mathrm{te}} = \sqrt{52 \sum_{i=1}^{\mathrm{s}} \left(w_{i,\mathrm{p}} - w_{i,\mathrm{b}}\right)^2 \varepsilon_{i,\mathrm{w}}^2} \tag{2.36}$$

The marginal volatility of a security, the measure of the sensitivity of the portfolio volatility is relative to the change in the specific security weight. We must know the relative contribution of each security to the risk of the portfolio.

$$\partial_{\mathrm{s}} = \frac{\partial \sigma_{\mathrm{p}}}{\partial w_{\mathrm{s}}} \tag{2.37}$$

where $\partial_{\mathrm{s}} =$ marginal risk of security s.

$$\partial_s = \beta_{s.\mathrm{p}} \sigma_{\mathrm{p}} \tag{2.38}$$

The APT marginal security volatility may be written as

$$\partial_{\mathrm{s}} = \frac{\sqrt{52 \left(\sum_{C=1}^{\mathrm{e}} \beta_{\mathrm{s,c}} \left(\sum_{i=1}^{\mathrm{s}} w_i \beta_{i,\mathrm{c}} \right) + w_{\mathrm{s}} \varepsilon_{\mathrm{s,w}}^2 \right)}}{\sqrt{52 \left(\sum_{C=1}^{\mathrm{c}} \left(\sum_{i=1}^{\mathrm{s}} w_i \beta_{i,\mathrm{c}} \right)^2 + \sum_{i=1}^{\mathrm{s}} w_i^2 \varepsilon_{i,\mathrm{w}}^2 \right)}} \tag{2.39}$$

Thus, the marginal security systematic volatility is the partial derivative of the systematic volatility of the portfolio relative to the security weight. In the King's English, the marginal tracking error measures the sensitivity of the tracking error relative to the marginal change in the security active weight.

$$\partial_{\mathrm{s},\beta} = \frac{52 \sum_{C=1}^{\mathrm{c}} \beta_{\mathrm{s,c}} \left(\sum_{i=1}^{\mathrm{s}} w_i \beta_{i,\mathrm{c}} \right)}{\sqrt{52 \sum_{C=1}^{\mathrm{c}} \left(\sum_{i=1}^{\mathrm{s}} w_i \beta_{i,\mathrm{c}} \right)^2}} \tag{2.40}$$

The marginal tracking error of a security is the partial derivative of the portfolio relative to the security active weight.

$$\partial_{\mathrm{s,te}} = \frac{52 \left(\sum_{C=1}^{\mathrm{c}} \beta_{\mathrm{s,c}} \left(\sum_{i=1}^{\mathrm{s}} \left(w_{i,\mathrm{p}} - w_{i,\mathrm{b}}\right) \beta_{i,\mathrm{c}} + w_{s,\mathrm{p}} - w_{s,\mathrm{b}} \right) \varepsilon_{\mathrm{s,w}}^2 \right)}{\sqrt{52 \left(\sum_{C=1}^{\mathrm{c}} \left(\sum_{i=1}^{\mathrm{s}} \left(w_{i,\mathrm{p}} - w_{i,\mathrm{b}}\right) \beta_{i,\mathrm{c}} \right)^2 + \sum_{i=1}^{\mathrm{s}} \left(w_{i,\mathrm{p}} - w_{i,\mathrm{b}}\right)^2 \varepsilon_{i,\mathrm{w}}^2 \right)}} \tag{2.41}$$

If a position taken in a security leads to an increase in the portfolio's volatility, then the security is said to create a positive contribution to risk. A negative contribution to risk occurs when a security reduces the portfolio volatility such as a long position on a security with a negative beta or a short position on a security with a positive beta. Obviously the contribution to risk depends upon the security weight and the security's beta to the overall portfolio. The security contribution to tracking error, ξ_{s}, reflects the security's contribution to the tracking error of a portfolio considering the security return that is undiversified at the active portfolio level.

$$\xi_{\mathrm{s}} = \left(w_{\mathrm{s,p}} - w_{\mathrm{s,b}}\right) \frac{52\left(\sum_{C=1}^{\mathrm{c}} \beta_{\mathrm{s,c}} \left(\sum_{i=1}^{\mathrm{s}} \left(w_{i,\mathrm{p}} - w_{i,\mathrm{b}}\right) \beta_{i,\mathrm{c}} + \left(w_{i,\mathrm{p}} - w_{\mathrm{s,b}}\right)\right) \varepsilon_{\mathrm{s,w}}^2\right)}{\sqrt{52\left(\sum_{C=1}^{\mathrm{c}} \left(\sum_{i=1}^{\mathrm{s}} \left(w_{i,\mathrm{p}} - w_{i,\mathrm{b}}\right) \beta_{i,\mathrm{c}}\right)^2 + \sum_{i=1}^{\mathrm{s}} \left(w_{i,\mathrm{p}} - w_{i,\mathrm{b}}\right)^2 \varepsilon_{i,\mathrm{w}}^2\right)}} \tag{2.42}$$

The portfolio's forecast beta, β, may be written:

$$\beta = \frac{\left(\sum_{C=1}^{\mathrm{c}} \left(\sum_{i=1}^{\mathrm{s}} w_{i,\mathrm{p}} \beta_{i\mathrm{c}}\right) \left(\sum_{i=1}^{\mathrm{s}} w_{i,\mathrm{b}} \beta_{i\mathrm{c}}\right) + \sum_{i=1}^{\mathrm{s}} w_{i,\mathrm{p}} w_{i,\mathrm{b}} \varepsilon_{i,\mathrm{w}}^2\right)}{\left(\sum_{C=1}^{\mathrm{c}} \left(\sum_{i=1}^{\mathrm{s}} w_{i,\mathrm{b}} \beta_{i,\mathrm{c}}\right)^2 + \sum_{i=1}^{\mathrm{s}} w_{i,\mathrm{b}}^2 \varepsilon_{i,\mathrm{w}}^2\right)} \tag{2.43}$$

The portfolio Value-at-Risk (VaR) is the expected maximum loss that a portfolio could produce over 1 year.

The APT measure of portfolio risk estimating the magnitude that the portfolio return may deviate from the benchmark return over 1 year is referred to as TaR, or "Tracking-at-Risk"™.

$$T_\sigma = \sqrt{\left(\frac{1}{\sqrt{1-x}} \sigma_{\mathrm{s}}\right)^2 + \left(\sqrt{2} \operatorname{erf}^{-1}(x) \sigma_\varepsilon\right)^2} \tag{2.44}$$

where T_σ = TaR™, x = Desired confidence level of TaR™, σ_{s} = Portfolio systematic tracking error, $\operatorname{erf}^{-1}(x)$ = inverse error function, and σ_ε = Portfolio specific tracking error.

Blin and Bender during the 1987–1997 period, developed an APT software system which estimated a 20 (factor) beta model of covariances based on 2.5 years of weekly stock returns data. The Blin and Bender Arbitrage Pricing Theory (APT) model followed the Ross factor modeling theory, but Blin and Bender estimated betas from 20 to 24 orthogonal factors. Blin and Bender never sought to identify their factors with economic variables.

The Guerard et al. (2009) contributed chapter in his volume compares the Blin and Bender TaR trade-off curve with an alternative Enhanced Index Tracking, denoted EIT, where Markowitz creates a maximum allowable deviation of the absolute value of the portfolio weight and the index (target) weight, such as a 2% maximum allowable deviation. The EIT formulation generally tracks an index far closer than the mean–variance framework. Why? Excess returns are found more often in smaller stocks than larger stocks, as we noted in the CTEF Barra example. If a manager produces a mean–variance portfolio with a 4% upper bound then several stocks not in the index could have a 4% position, whereas the maximum position might be a 2% position in the Markowitz EIT formulation.[6]

In the early 1990s, with the Markowitz Global Portfolio Research team at Daiwa Securities, Harry ran an equity-only research group that modeled global securities, Japanese equities in particular. Guerard et al. (1993) published a study that compared mean–variance portfolio construction with an EIT portfolio construction techniques using the first-section, non-financial stocks of the Tokyo Stock Exchange for the January 1974–December 1990 period. The paper concluded that an asset manager can create portfolios that maximize expected returns and minimize expected portfolio tracking error, rather than just portfolio total risks. Guerard et al. (1993) measured risk vs. a tracking error concept, the monthly semi-variance of the benchmark, the TOPIX. We assumed quarterly re-optimization of a robust regression-weighted composite model of expected returns (described in Chap. 3), 10% turnover constraints, 2% upper bounds on security weights, and 2% (each way) transactions costs. The domination of EIT over mean–variance is shown in Fig. 2.4. Each point on the trade-off curves, or estimated efficient frontiers, represents a percentile-point between maximizing pure expected return and minimizing risk. The EIT technique produces a higher geometric mean relative to the tracking error when compared with mean–variance analysis. Using a different measure of risk, the Sharpe Ratio (portfolio excess return relative to the portfolio standard deviation [Sharpe 1966]), mean–variance portfolio construction techniques were favored over EIT analysis.[7]

One sees that an efficient frontier can be constructed using mean–variance analysis, shown in Chap. 1, or with an enhanced index tracking techniques, as illustrated below. The EIT trade-off curve dominates the MV trade-off curve on the basis of the portfolio tracking error, measured by the portfolio monthly semi-deviation with the relevant benchmark or index, the TOPIX index, in this example.

[6] Markowitz discusses the MFM formulation in his second monograph, *Mean-Variance Analysis in Portfolio Choice and Capital Markets*, New Hope, PA: Frank J. Fabozzi Associates, 2000), Chap. 3, p. 50.

[7] The optimal Sharpe Ratio on the mean-variance efficient frontier occurred when one went 90% along the return-to-risk frontier. Markowitz et al. (1993) and Guerard et al. (1993) referred to the distance along the efficient frontier as the "Pick Parameter," denoted PPar. The MV Sharpe ratio was 0.90 at the PPar $=$ 0.90 whereas the EIT Sharpe Ratio was 0.81 at the PPar $=$ 0.95. The use of the EIT allows one to get a smaller tracking error than the MV analysis and move toward a higher PPar than the MV analysis, if one is concerned with tracking error. Rudd and Clasing (1982) and Grinold and Kahn (2000) maximize the Information Ratio, the portfolio excess return relative to tracking error. The PPar is analogous to the Sharpe lambda discussed earlier in this chapter and the APT-based lambda used in Guerard et al. (2008).

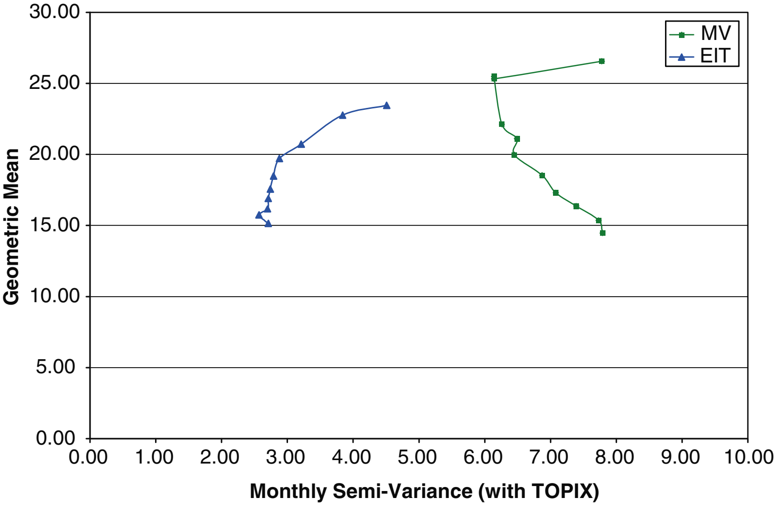


Fig. 2.4 Mean–variance (MV) and enhanced index tracking (EIT) analysis, first section stocks of the Tokyo Stock Exchange, 1974–1990

Asset managers and institutional investors have worked together to manage active, or tracking error, risk. Fabbozi et al. (2002) stated that one needs to construct the implicit relationship between active manager performance and the portfolio deviation from their benchmark. Index managers must track the index such that their expected alpha is zero with no tracking error. The enhanced index manager is expected to generate positive, but relatively small alpha, 1%, with relatively small tracking error, 1.75%, in the Fabozzi et al. (2002) example. Core managers earn slightly higher alphas with larger tracking errors. Active managers, performing stock selection, earn the largest alpha but have the largest tracking errors.

Who should determine whether one uses mean–variance analysis or an enhanced index tracking analysis? Perhaps one must listen to one's client and ask what is more important, maximizing expected terminal wealth, and enhance the geometric mean and the Sharpe Ratio, leading to higher lambda values, or maximizing the Information Ratio, which minimizing tracking error relative to portfolio annualized residual return. We will discuss a numerical example of creating another efficient frontier in Chap. 3.

2.5 Optimal Portfolio Weighting: The Use of Beta in Elton, Gruber, Brown, and Goetzmann

An alternative or model manner to solve for optimal security portfolio weighting has been developed by Elton et al. (1978). Optimal portfolio construction can be derived using the market model and its estimated beta, as it was for the portfolio standard deviation. Defining

$$s_{ij} = b_i b_j s_{\mathrm{M}}^2 \tag{2.45}$$

Elton, Gruber, and Padberg write the portfolio standard deviation as

$$\sigma_{\mathrm{p}} = \left[\sum w_i^2 \beta_i^2 \sigma_{\mathrm{M}}^2 + \sum_{i-1}^{N} \sum_{\substack{j=1 \\ i \neq j}}^{N} w_i w_j \beta_i \beta_j \sigma_{\mathrm{M}}^2 + \sum_{i=1}^{N} x_i^2 \sigma_{ei}^2 \right]^{1/2} \tag{2.46}$$

$$\theta = \frac{\sum_{i=1}^{N} w_i \left(E(R_i) - R_{\mathrm{F}}\right)}{\left[\sum_{i=1}^{N} w_i^2 \beta_i^2 \sigma_{\mathrm{M}}^2 + \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \beta_i \beta_j \sigma_{\mathrm{M}}^2 + \sum{}_{w_i}^{2} \sigma_{e_i}^2 \right]^{1/2}} \tag{2.47}$$

When an asset manager or investor seeks to maximize θ, he or she seeks to maximize portfolio expected returns to portfolio total risk. Total risk is now re-written in terms of security betas. Theta, as defined in (2.47), is the Sharpe Ratio recast in Beta terms.

$$\frac{\partial \theta}{\partial x_i} = E(R_i) - R_{\mathrm{F}} - \sum_{\substack{i=1 \\ j \neq i}}^{N} w_i \left(E(R_i) - R_{\mathrm{F}}\right) \left[w_i\, \beta_i^2 \sigma_{\mathrm{M}}^2 + \beta_i \sum_{j=1}^{N} w_j \beta_j \sigma_{\mathrm{M}}^2 + w_i \sigma_{e_i}^2 \right] = 0 \tag{2.48}$$

Optimal security weights are determined by taking derivatives of total risk with respect to security weights.

$$z_i = \left(\frac{E(R_{\mathrm{p}}) - R_{\mathrm{F}}}{\sigma_{\mathrm{p}}^2} \right) w_i$$

Then Elton, Gruber, and Padberg can re-write the optimal portfolio scaling factor as

$$z_i = \frac{E(R_i) - R_{\mathrm{F}}}{\sigma_{e_i}^2} - \left[\sigma_{\mathrm{M}}^2 \frac{\sum_{j=1}^{N} \frac{E(R_j) - R_{\mathrm{F}}}{\sigma_{e_i}^2} \beta_j}{1 + \sigma_{\mathrm{M}}^2 \sum_{j=1}^{N} \frac{\beta_j^2}{\sigma_{ej}^2}} \right] \frac{\beta_i}{\sigma_{ei}^2} \tag{2.49}$$

Finally, Elton, Gruber, and Padberg solve for the optimal portfolio weighting scheme, where

$$z_i = \frac{\beta_i}{\sigma_{e_i}^2} \left(\frac{\bar{R}_i - R_F}{\beta_i} - C^* \right) \tag{2.50}$$

C^* denotes the portfolio-maximizing excess return relative to beta ratio

$$w_i = \frac{z_i}{\sum_{j=1}^{N} z_j}$$

Jacobs and Levy (2009) and Jacobs et al. (2006) show the algorithmic equivalence between the Elton, Gruber, and Padberg formulation and the Markowitz Critical Line Algorithm in their chapter in this volume.

2.6 Conclusions and Summary

This chapter addresses several aspects of estimating betas and the estimation and usefulness of multi-factor models of risk. In the early 1970s, Barr Rosenberg, James Farrell, Bernell Stone, Steve Ross, Richard Roll, Phoebus Dhrymes, Irwin Friend, Mustafa Gultekin, and John Blin among others developed and estimated multi-factor risk models. Some models have survived the test of time. Still fewer became commercially viable within the investment management community. Markowitz analysis has long been aware of different procedures of estimating risk.

Appendix: USE3 Descriptor Definitions

This Appendix gives the definitions of the descriptors which underlie the risk indices in US-E3. The method of combining these descriptors into risk indices is proprietary to MSCI/Barra.

1. *Volatility* is composed of variables including the historic beta, the daily standard deviation, the logarithm of the stock price, the range of the stock return relative to the risk-free rate, the options pricing model standard deviation, and the serial dependence of market model residuals.
2. *Momentum* is composed of a cumulative 12-month relative strength variable and the historic alpha from the 60-month regression of the security excess return on the S&P 500 excess return.
3. *Size* is the log of the security market capitalization.
4. *Size Nonlinearity* is the cube of the log of the security market capitalization.
5. *Trading Activity* is composed of annualized share turnover of the past 5 years, 12 months, quarter, and month, and the ratio of share turnover to security residual variance.
6. *Growth* is composed of the growth in total assets, 5-year growth in earnings per share, recent earnings growth, dividend payout ratio, change in financial leverage, and analyst-predicted earnings growth.
7. *Earnings Yield* is composed of consensus analyst-predicted earnings to price and the historic earnings to price ratios.
8. *Value* is measured by the book to price ratio.

9. *Earnings Variability* is composed of the coefficient of variation in 5-year earnings, the variability of cash flow, and the variability of analysts' forecasts of earnings to price.
10. *Leverage* is composed of market and book value leverage, and the senior debt ranking.
11. *Currency Sensitivity* is composed of the relationship between the excess return on the stock and the excess return on the S&P 500 Index. These regression residual returns are regressed against the contemporaneous and lagged returns on a basket of foreign currencies.
12. *Dividend Yield* is the Barra-predicted dividend yield.
13. *Non-Estimation Universe Indicator* is a dummy variable which is set equal to zero if the company is in the Barra estimation universe and equal to one if the company is outside the Barra estimation universe.

References

APT. Inc. 2005. *Analytics Guide*.

Black, F., M.C. Jensen, and M. Scholes. 1972. "The Capital Asset Pricing Model: Some Empirical Tests." In M. Jensen (ed.), *Studies in the Theory of Capital Markets*. New York: Praeger.

Blin, J., M.S. Bender, and J.B. Guerard Jr. 1997. "Earnings Forecasts, Revisions and Momentum in the Estimation of Efficient Market-Neutral Japanese and U.S. Portfolios." In A. Chen (ed.), *Research in Finance* 15.

Chen, N-F., R. Roll, and S. Ross. 1986. "Economic Forces and the Stock Market." *Journal of Business* 59, 383–403.

Cohen, K.J. and J.A. Pogue. 1967. "An empirical evaluation of alternative portfolio-selection models." *Journal of Business* 40, 166–193.

Conner, G. and R.A. Korajczyk. 1995. "The Arbitrary Theory and Multifactor Models of Asset Returns." In R.A. Jarrow, V. Malsimovic, and W.T. Ziemba (eds.), *Handbooks in Operations Research and Management Science: Finance* 9, 87–144.

Dhrymes, P.J., I. Friend, and N.B. Gultekin. 1984. "A Critical Re-examination of the Empirical Evidence on the Arbitrage Pricing Theory." *Journal of Finance* 39, 323–346.

Dhrymes, P.J., I. Friend, M.N. Gültekin, and N.B. Gültekin. 1985. "New Tests of the APT and Their Implications." *Journal of Finance* 40, 659–674.

Elton, E.J. and M.J. Gruber. 1970. "Homogeneous Groups and the Testing of Economic Hypothesis." *Journal of Financial and Quantitative Analysis* 5, 581–602.

Elton, E.J., M.J. Gruber, S. Brown, and W. Goetzmann. 2007. Modern Portfolio Theory and Investment Analysis. New York: Wiley. Eighth Edition.

Elton, E.J., M.J. Gruber, and M. Padberg. 1978. "Simple Criteria for Optimal Portfolio Selection: Tracing out the Efficient Frontier." *Journal of Finance* 33, 296–302.

Elton, E.J., M.J. Gruber, and M.W. Padberg. 1979. "Simple Criteria for Optimal Portfolio Selection: The Multi-Index Case." In E.J. Elton and M.J. Gruber (eds.), *Portfolio Theory, 25 Years After: Essays in Honor of Harry Markowitz*. Amsterdam: North-Holland.

Fabbozi, F., F. Gupta, and H.M. Markowitz. 2002. "The Legacy of Modern Portfolio Theory." *Journal of Investing*, Fall 2002, 7–22.

Fama, E.F. and J.D. MacBeth. 1973. "Risk, Return, and Equilibrium: Empirical Tests." *Journal of Political Economy* 81, 607–636.

Farrell, J.L., Jr. 1974. "Analyzing Covariance of Returns to Determine Homogeneous Stock Groupings." *Journal of Business* 47, 186–207.

Farrell, J.L. Jr. 1997. *Portfolio Management: Theory and Applications*. New York McGraw-Hill/Irwin.

Grinhold, R and R. Kahn. 1999. *Active Portfolio Management*. New York:. McGraw-Hill/Irwin.

Guerard J.B, Jr and A. Mark. 2003. "The Optimization of Efficient Portfolios: The Case for a Quadratic R&D Term." In A. Chen (ed.), *Research in Finance* 20, 213–247.

Guerard, J.B., Jr. M. Gultekin, and B.K. Stone. 1997. "The Role of Fundamental Data and Analysts' Earnings Breadth, Forecasts, and Revisions in the Creation of Efficient Portfolios." In A. Chen (ed.), *Research in Finance* 15.

Guerard, J.B., Jr., M. Takano, and Y. Yamane. 1993. "The Development of Efficient Portfolios in Japan with Particular Emphasis on Sales and Earnings Forecasting." *Annals of Operations Research* 45, 91–108.

Jacobs, B.I. and K.N. Levy. 2009. "Reflections on Portfolio Insurance, Portfolio Theory, and Market Simulation with Harry Markowitz." This volume.

King, B.F. 1966. "Market and Industry Factors in Stock Price Behavior." *Journal of Business* 39, 139–191.

Markowitz, H.M. 2000. *Mean–Variance Analysis in Portfolio Choice and Capital Markets*. New Hope, PA: Frank J. Fabozzi Associates.

Mossin, J. 1966. "Equilibrium in a Capital Asset Market." *Econometrica* 34, 768–783.

Mossin, J. 1973. *Theory of Financial Markets*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Ramnath, S., S. Rock, and P. Shane. 2008. "The Financial Analyst Forecasting Literature: A Taxonomy with Suggestions for Further Research." *International Journal of Forecasting* 24, 34–75.

Roll, R. 1969. "Bias in Fitting the Sharpe Model to Time Series Data." *The Journal of Financial and Quantitative Analysis* 4, 271–289.

Roll, R. 1977. "A Critique of the Asset Pricing Theory's Tests, Part I: On Pat and Potential Testability of the Theory." *Journal of Financial Economics* 4, 129–176.

Rosenberg, B. 1974. "Extra-Market Components of Covariance in Security Returns." *Journal of Financial and Quantitative Analysis* 9, 263–274.

Rosenberg, B., M. Hoaglet, V. Marathe and W. McKibben. 1975. "Components of Covariance in Security Returns." Working paper no. 13, Research Program in Finance. Institute of Business and Economic Research. University of California. Berkeley.

Rosenberg, B. and W. McKibben. 1973. "The Prediction of Systematic and Unsystematic Risk in Common Stocks." *Journal of Financial and Quantitative Analysis* 8, 317–333.

Rosenberg, B. and V. Marathe. 1979. "Tests of Capital Asset Pricing Hypotheses." In H. Levy (ed.), *Research in Finance* 1.

Rosenberg, B. and A. Rudd. 1977. "The Yield/Beta/Residual Risk Tradeoff." Working paper no. 66. Research Program in Finance. Institute of Business and Economic Research. University of California, Berkeley.

Ross, S.A. 1976. "The Arbitrage Theory of Capital Asset Pricing." *Journal of Economic Theory* 13, 341–360.

Ross, S.A. and R. Roll. 1980. "An Empirical Investigation of the Arbitrage Pricing Theory." *Journal of Finance* 35, 1071–1103.

Rudd, A. and B. Rosenberg. 1980. "The 'Market Model' in Investment Management." *Journal of Finance* 35, 597–607.

Rudd, A, and B. Rosenberg. 1979. "Realistic Portfolio Optimization." In E. Elton and M. Gruber (eds.), *Portfolio Theory, 25 Years After*. Amsterdam: North-Holland.

Sharpe, W.F. 1963. "A Simplified Model for Portfolio Analysis." *Management Science* 9, 277–293.

Sharpe, W.F. 1964. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *Journal of Finance* 19, 425–442.

Sharpe, W.F. 1966. "Mutual Fund Performance." *Journal of Business* 39, 119–138.

Sharpe, W.F. 1970. *Portfolio Theory and Capital Markets*. New York: McGraw-Hill.

Sharpe, W.F. 1971. "A Linear Programming Approximation for the General Portfolio Analysis Problem." *Journal of Financial and Quantitative Analysis*. 6, 1263–1275.

Sharpe, W.F. 1971. "Mean-Absolute-Deviation Characteristic Lines for Securities and Portfolios." *Management Science* 18, B1–B13.

Stone, B.K. 1970. *Risk, Return, and Equilibrium: A General Single-Period Theory of Asset Selection and Capital Market Equilibrium*. Cambridge, MA: MIT Press.

Stone, B.K. 1973. "A Linear Programming Formulation of the General Portfolio Selection Problem." *Journal of Financial and Quantitative Analysis* 8, 621–636.

Stone, B.K. 1974. "Systematic Interest-Rate Risk in a Two-Index Model of Returns." *Journal of Financial and Quantitative Analysis* 9, 709–721.

Stone, B.K., and J.B. Guerard Jr. 2009. "Methodologies for Isolating and Assessing Potential Portfolio Performance of Stock Return Forecast Models with an Illustration." This volume.

Stone, B.K., J.B. Guerard Jr., M. Gultekin, and G. Adams. 2002. "Socially Responsible Investment Screening." Working Paper. Marriott School of Management. Brigham Young University.

Von Hohenbalken, B. 1975. "A Finite Algorithm to Maximize Certain Pseudoconcave Functions on Polytopes." *Mathematical Programming* 8, 189–206.

Chapter 3
Markowitz Applications in the 1990s and the New Century: Data Mining Corrections and the 130/30

John B. Guerard, Jr.

In Chap. 1, we introduced the reader to Harry Markowitz and portfolio theory, as advanced in the *Portfolio Selection* (1959) monograph. In Chap. 2, we introduced the reader to multifactor risk models and the estimation of efficient frontiers, model testing, and mean variance, and enhanced index-tracking models. In this chapter, we introduce the reader to four important issues that Markowitz has worked on during the 1990–2007 period. First, Professor Markowitz and the Global Portfolio Research Department (GPRD) at Daiwa Securities worked on fundamental variables, expected return forecasting, and portfolio construction. Second, Professor Markowitz and Ganlin Xu of GPRD developed and estimated a Data Mining Corrections test to assess whether the excess portfolio returns were statistically significant. Third, Harry worked with Bruce Jacobs and Ken Levy to develop the Jacobs Levy Markowitz Financial Simulator of financial market behavior. Fourth, Harry worked with Jacobs and Levy on portfolio optimization including realistic short positions. Markowitz's 1959 monograph illustrated expected return, using average values of historical returns. This chapter examines how fundamental data such as earnings, book value, cash flow, dividends, net current assets, and price momentum, and the expectational data such as analysts' forecasts, forecast revisions, and direction of revisions can be used to predict returns. Markowitz and his colleagues have used two frameworks in particular during the past 15 years: first, creating a Data Mining Corrections tests to examine the statistical significance of excess returns and estimate how much of the excess returns might be expected to be realized in the future; and second, identifying securities expected to outperform the benchmark on the long-only side and securities expected to underperform on the short side. A portfolio that is invested 130% on the long side and 30% on the short side, known as a "130/30 portfolio", allows an investment manager more opportunities than traditional, long-only portfolios to exploit market inefficiencies.

J.B. Guerard, Jr. (✉)
McKinley Capital Management, LLC, Anchorage, AK, USA
e-mail: jguerard@mckinleycapital.com

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_3,
© Springer Science+Business Media, LLC 2010

3.1 Expected Returns Vs. Historical Mean Returns: The Creation and Testing of Fundamental and Expectational Data and their Role in Portfolio Construction

Expected returns on assets are not completely explained by using only historical means (and standard deviations). One can estimate models of expected returns, using expectational data and reported financial data. There are several approaches to security valuation and the creation of expected returns. Graham et al. (1934) recommended that stocks be purchased on the basis of the price–earnings (P/E) ratio. They suggested that no stock should be purchased if its P/E ratio exceeded 1.5 times the P/E multiple of the market. Thus, the P/E criteria were established. It is interesting that Graham and Dodd put forth the low P/E model at the height of the Great Depression. Basu (1977) reported evidence supporting the low P/E model. Academicians often prefer to test the low P/E approach by testing its reciprocal, the "high EP" approach. The high-EP approach specifically addresses the issue of negative earnings per share, which can confuse the low-P/E test. Hawawini and Keim (1995) found statistical support for the high-EP variable of NYSE and AMEX stocks from April 1962 to December 1989. At a minimum, Graham and Dodd also advocated the calculation of a security's net current asset value, NCAV, defined as its current assets less all liabilities. A security should be purchased if its net current value exceeded its current stock price. The price-to-book (PB) ratio should be calculated, but not used as a measure for stock selection. Academicians test the book-to-price ratio, BP, as opposed to the PB ratio because of negative book values and discontinuous ratio values.

In addition to the extensive literature on the impact of individual value ratios on the cross-section of stock returns, there is also a massive literature on the use of value ratios to predict stock special situations such as mergers or financial misreporting. We concentrate much of our review to a very specialized subset of value-focused, stock return forecasting, namely one-step regression-estimated composites of three or more value ratios, especially those that go beyond using just one or two of the standard value ratios (DP, EP, and/or BP) to include the cash-price ratio (CP) and/or the sales-price ratio (SP).[1] The major papers on combination of value ratios to predict stock returns that include at least CP and/or SP include Chan et al. (1991), Bloch et al. (1993), Lakonishok et al. (1994), and Guerard et al. (1997). We review these four papers in some detail.

Chan et al. (1991) used seemingly unrelated regression (SUR) to model CAPM monthly excess returns as functions of CAPM excess returns of the value-weighted

[1] Jacobs and Levy (1988) presented an extensive survey of the anomaly literature. Before 1998, 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 "pure" (i.e., underlying) anomalous effects from what they called the "naïve" 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.

or equal-weighted market index return; EP, BP, and CP; and size as measured by the natural logarithm of market capitalization (LS).[2] When fractile portfolios were constructed by sorting on the EP ratio, the highest EP quintile portfolio outperformed the lowest EP quintile portfolio, although the EP effect was not statistically significant. The highest BP stocks outperformed the lowest BP stocks. The portfolios composed (sorted) of the highest BP and CP outperformed the portfolios composed of the lowest BP and CP stocks. In the authors' multiple regressions, the size and book-to-market variables were positive and statistically significant. The EP coefficient was negative and statistically significant at the 10% level. Thus, no statistically significant support was found for the Graham and Dodd low-PE approach. In the monthly univariate SUR analysis, with each month variable being deflated by an annual (June) cross-sectional mean, Chan et al. (1991) found that the EP coefficient was negative (but not statistically significant), the size coefficient was negative (but not statistically significant), the book-to-market coefficient was positive and statistically significant, and the cash flow coefficient was positive and statistically significant. In their multiple regressions, Chan et al. (1991) report that BP and CP variables were positive and statistically significant but EP was not significant. Applying an adaptation of the Fama and MacBeth (1973) time series of portfolio cross-sections to the Japanese market produced negative and statistically significant coefficients on EP and size but positive and statistically significant coefficients for the BP and CP variables. Chan et al. (1991, p. 1760) summarized their findings: "The performance of the book-to-market ratio is especially noteworthy; this variable is the most important of the four variables investigated."

Let R denote stock (total) return for a typical stock. Bloch et al. (1993) estimate the following regression equation to assess empirically the relative explanatory power of each of the eight value ratios in the equation:

$$R = w_0 + w_1\mathrm{EP} + w_2\mathrm{BP} + w_3\mathrm{CP} + w_4\mathrm{SP} + w_5\mathrm{REP} + w_6\mathrm{RBP} + w_7\mathrm{RCP} + w_8\mathrm{RSP} + e_t. \tag{3.1}$$

Given concerns about both outlier distortion and multicollinearity, Bloch et al. (1993) tested the relative explanatory and predictive merits of alternative regression estimation procedures: OLS, robust using the Beaton and Tukey (1974) bi-square criterion to mitigate the impact of outliers, latent root to address the issue of multicollinearity (see Gunst et al. 1976), and weighted latent root, denoted WLRR, a combination of robust and latent root.

Bloch et al. (1993) used the estimated regression coefficients to construct a rolling horizon return forecast. The predicted returns and predictions of risk parameters were used as input to a mean-variance optimizer (see Markowitz 2000) to create mean-variance efficient portfolios in both Japan (first section, nonfinancial Tokyo Stock Exchange common stocks, January 1975–December 1990) and the United States (the 1,000 largest market-capitalized common stocks, November 1975–December 1990).[3] Bloch et al. (1993) found that Markowitz (2000) mean-variance

[2] Chan et al. (1991) define cash as the sum of earnings and depreciation without any explicit correction for other noncash revenue or expenses.

[3] The use of non-financial stocks led to a customized index for the Markowitz Global Portfolio Research Group (GPRD) analysis. The Chan et al. and an initial Guerard presentation occurred

efficient portfolios using the higher EP values in Japan underperformed the universal benchmark, whereas BP, CP, and SP (sales-to-price or sales yield) variables outperformed the universal benchmark. For the United States optimized portfolios using BP, CP, SP, and EP variables, the portfolios outperformed the United States S&P 500 index, giving support to the Graham–Dodd concept of using the relative rankings of value-focused fundamental ratios to select stocks.[4]

Bloch et al. (1993) used relative ratios as well as current ratio values in identifying mis-priced stocks. An investor might want to purchase not just a low-P/E stock, but a low-P/E stock when the P/E is at a relatively low value compared to its historical value, in this case a low relative to its average over the last 5 years. Bloch et al. (1993) reported several results. First, for both Japan and the United States' financial markets, they compared OLS and WLLR techniques, inputting expected returns forecasts produced by each method into a mean-variance optimizer. The WLRR-constructed, composite model portfolio produced higher Sharpe ratios and geometric means than the OLS-constructed, composite variable portfolio in both Japan and the United States, indicating that controlling for both outliers and multicollinearity is important in using regression-estimated composite forecasts.

Second, Bloch et al. quantified survivor bias and found it was not statistically significant in Japan and the US for the period tested. Third, they investigated period-to-period portfolio revision and found that tighter turnover and rebalancing triggers led to higher portfolio returns for value-based strategies. Finally, Markowitz and Xu (1994) developed a test for data mining. In addition to testing the hypothesis of data mining, the test can be used to estimate and assess the expected difference between the best test model and the average of simulated policies.

In a thorough assessment of value versus growth in the United States, Lakonishok et al. (1994) examined the intersection of Compustat and CRSP databases for annual portfolios for NYSE and AMEX common stocks, April 1963–April 1990. Their value measures were three current value ratios: EP, BP, and CP. Their growth measure was the 5-year average annual sales growth (GS). They performed three types of tests: a univariate ranking into annual decile portfolios for each of the four variables, bivariate rankings on CP (value) and GS (growth, glamor), and finally a multivariate regression adaptation of the Fama and MacBeth (1973) time-series pooling of cross-sectional regressions.

---

in September 1991 at the Berkeley Program in Finance, Santa Barbara, on Fundamental Analysis. Bill Ziemba presented a very interesting study comparing US and Japanese fundamental strategies at the same Berkeley Program meeting. Markowitz refers to this meeting in his Nobel Prize lecture (1991).

[4] One finds the price/earnings, price/book, and price/sales listed among the accounting anomalies in Levy (1999), p. 434. Levy also discusses the dividend yield as a (positive) stock anomaly. Malkiel (1996) cites evidence in support of buying low P/E, low P/B, and high D/P (dividend yield) stocks for outperformance, provided the low P/E stocks have modest growth prospects (pp. 204–210). Malkiel speaks of a "double bonus"; that is, if growth occurs, earnings increase and the price-to-earnings multiple may increase, further driving up the price. Of course, should growth fail to occur, both earnings and the P/E multiple may fall.

Lakonishok et al. (1994) initially used a univariate sort on each variable to form 10 annual fractile portfolios for each year from April 1968 through April 1989. The average high-minus-low spreads for the univariate decile rankings were 10.0, 11.0, 7.6, and −2.4% (on a size-related basis) for BP, CP, EP, and GS, respectively. In the two-way, CP–GS bivariate sort, the extreme value portfolio (highest CP, lowest GS) returned 22.1%, whereas the glamor portfolio (highest GS, lowest CP) returned only 11.4%, a difference of 10.7%. Moreover, low CP stocks with low historical sales growth underperformed low CP stocks with high past sales growth by over 4.5% points.

Lakonishok et al. (1994) used the Fama–MacBeth methodology to construct portfolios and pool (average over time) a time series of 22 one-year, cross-sectional, univariate regressions for each of the 22 years in their study period. The univariate regression coefficient for SG was significantly negative. The EP, BP, and CP coefficients were all significantly positive. When Lakonishok, Shleifer, and Vishny performed a multivariate regression using all four variables, they found significantly positive coefficients for BP and EP (but not CP) and significantly negative coefficients for SG. Overall, Lakonishok et al. (1994) concluded that buying out-of-favor value stocks outperformed growth (glamor) over the April 1968–April 1990 period, that future growth was difficult to predict from the past growth alone and that the actual future growth of the glamor stocks was much lower than the past growth relative to the growth of value stocks, and that the value strategies ex post were not significantly riskier than growth (glamor) strategies.

Guerard et al. (1997) studied the intersection of Compustat, CRSP, and I/B/E/S (the Institutional Brokerage Estimation Service database) databases. This study was built on the fundamental forecasting work in Bloch et al. (1993) in two ways: adding to the Bloch et al.'s eight-variable regression equation, a growth measure, and then also adding three measures of analysts' forecasts and forecast revisions from the I/B/E/S database, namely consensus analysts' forecasts, forecast revisions, and the direction (net up or down) of the forecast revisions.

In quarterly weighted, latent-root regressions, the growth variable averaged a relative weight of 33% whereas the average relative weighting of the eight value variables averaged almost 67%. This result complements that of Lakonishok et al. (1994) in showing that the rank-ordered portfolio returns have both a significant value and growth components.

Adding I/B/E/S variables to the eight value ratios produced more than 2.5% of additional annualized return. The finding of a significant, predictive performance value for the three I/B/E/S variables indicates that the analyst forecast information has value beyond purely statistical extrapolation of past value and growth measures. Possible reasons for the additional performance benefit could be that analysts' forecasts and forecast revisions reflect information in other return-pertinent variables, or discontinuities from past data, or serve as a quality screen on the otherwise out-of-favor stocks. The quality-screen idea would confirm Graham and Dodd's argument that value ratios should be used in the context of the many qualitative and quantitative factors that they argue are essential to informed investing. In terms of a relative predictive value, Guerard et al. (1997) found the EP, CP, SP, and RSP variables to be more important than the BP variable.

To test the risk-corrected performance value of the forecasts, Guerard et al. (1997) formed quarterly portfolios with risk being modeled via a four-factor, APT-based model (created using 5 years of past monthly data). The portfolios' quarterly returns averaged 6.18% before correcting for risks and transaction costs, with excess returns of 3.6% after correcting for risk, and 2.6% quarterly after subtracting 100 basis points to reflect an estimate of two-way transactions costs. The annualized, after-risk, after-cost return of 10.4% is both economically and statistically significant.

Fundamental variables such as cash flow and sales have been used in composite valuation models for security selection (Ziemba 1990, 1992). In addition to the income statement indicators of value, such as earnings, cash flow, and sales, many value-focused analysts also considered balance-sheet variables, especially the book-to-market ratio. The income statement measures are dividends, earnings, cash flow, and sales and the key balance sheet measure is common equity per share outstanding, or book value. Expected returns modeling has been analyzed with a regression model in which security returns are functions of fundamental stock data, such as earnings, book value, cash flow, and sales, relative to stock prices, and forecast earnings per share (Fama and French 1992, 1995; Bloch et al. 1993; Guerard et al. 1993; Ziemba 1992; Guerard et al. 1997). Hawawini and Keim (1995) found statistical support for the high cash flow-to-price (CP) and low price-to-book (P/B) variables of NYSE and AMEX stocks from April 1962 to December 1989.

Fama and French (1992, 1995) argue that the contrarian investment strategy is inconsistent with market data. Fama and French report results of a three-factor model that incorporates the market factor, size factor, and book-to-market ratio. In the Fama and French analysis of the CRSP database for the 1963–1994 period, the value-weighted market return exceeded the risk-free rate by approximately 5.2%, annually, whereas the smaller stocks outperformed the larger stocks by 3.2% annually, and the higher book-to-market (price) outperformed the smaller book-to-market stocks by 5.4%, annually. The Fama and French model may be written as:

$$R_j - R_{\mathrm{F}} = \beta_{\mathrm{M}} R_{\mathrm{M}} + \beta_{\mathrm{size}} R_{\mathrm{size}} + \beta_{\mathrm{BM}} R_{\mathrm{BM}}, \tag{3.2}$$

where

R_{j} = stock return;
β_{M} = market beta;
R_{M} = return on the market index less the risk-free rate;
β_{size} = size beta;
R_{size} = return on the size variable, defined as the return on small stocks less the return on large stocks;
β_{BM} = book-to-market beta;
and R_{BM} = returns on high book-to-market stocks less the returns on low book-to-market stocks.

The Fama and French three-factor model is often cited in the literature. The stock selection model introduced and estimated in the following section does not place a higher weight on the book-to-market (price) ratio.

In 1975, a database of earnings per share (eps) forecasts was created by Lynch, Jones, and Ryan, a New York brokerage firm, by collecting and publishing the consensus statistics of 1-year-ahead- and 2-year-ahead eps forecasts (Brown 1999). The database evolved to become known as the Institutional Brokerage Estimation Service (I/B/E/S) database. There is an extensive literature regarding the effectiveness of analysts' earnings forecasts, earnings revisions, earnings forecast variability, and breadth of earnings forecast revisions, summarized in Brown (1999), and Ramnath et al. (2008). The vast majority of the earnings forecasting literature in the Brown references find that the use earnings forecasts do not increase stockholder wealth, as specifically tested in Elton et al. (1981). Reported earnings follow a random walk with drift process, and analysts are rarely more accurate than a no-change model in forecasting earnings per share, eps (Cragg and Malkiel 1968, Guerard and Stone 1992). Analysts become more accurate as time passes during the year, and quarterly data are reported. Analyst revisions are statistically correlated with stockholder returns during the year (Hawkins et al. 1984; Arnott 1985). Wheeler (1994) developed and tested a strategy in which analyst forecast revision breadth, defined as the number of upward forecast revisions less the number of downward forecast revisions, divided by the total number of estimates, was the criteria for stock selection. Wheeler found statistically significant excess returns from the breadth strategy. A composite earnings variable, CTEF, is calculated using equally-weighted revisions, forecasts, and breadth of FY1 and FY2 forecasts, a variable put forth in Guerard (1997).

Ziemba (1990, 1992), and Guerard et al. (1997) employed annual fundamental Compustat variables, such as earnings, book value, cash flow, and sales, in addition to the composite earnings forecasting model in a regression model to identify the determinants of quarterly equity returns. The regression models used in the Guerard et al. (1997) studies employed the Beaton–Tukey robust regression procedure and latent-root regression techniques to address the issues of outliers and multicollinearity. Ziemba used capitalization-weighted regressions. Both sets of studies found statistical significance with expectation and reported fundamental data.

3.2 Modern Portfolio Theory and GPRD: An Example of Markowitz Analysis

In 1990, Harry Markowitz became the Head of the Global Portfolio Research Department (GPRD) at Daiwa Securities Trust. His department used fundamental data to create models for Japanese and US securities. The basic models tested included the earnings-to-price (EPR) strategy, which is the inverse of the low P/E strategy, the book value-to-price (BPR), the cash flow-to-price (CPR), and sales-to-price (SPR) ratios. These variables should be positively associated (correlated) with subsequent returns. The researchers tested single-variable and composite model strategies for Japan and the US over 1974–1990. The composite models could be created by combining variables using ordinary least squares (OLS), outlier-adjusted or robust regression (ROB), or weighted latent-root regression (WLRR) modeling, in which outliers and the high correlations among the variables are used in the estimation procedure. The reader is referred to Bloch et al. (1993) for a discussion of ROB and WLRR techniques.[5] The Markowitz group found that the use of the more advanced statistical techniques produced higher relative, out-of-sample portfolio geometric returns and Sharpe ratios. Statistic modeling is not just fun, but is consistent with maximizing portfolio returns. The quarterly estimated models outperformed the semiannual estimated models, though the underlying data was semiannual in Japan. The dependent variable in the composite model is total security quarterly returns and the independent variables are the EPR, BPR, CPR, and SPR variables.

3.2.1 Further Estimations of a Composite Equity Valuation Model

Here we discuss issues of databases and the inclusion of variables in composite models to identify undervalued securities. The database for this analysis is created by the use of all securities listed on the Compustat database, the I/B/E/S database, and the Center for Research in Security Prices (CRSP) database during the 1987–2001 period. The annual Compustat file contains some 399 data items from the company income statement, balance sheet, and cash flow statement during the 1950–2007 period. The I/B/E/S database contains all earnings forecasts made during the 1976–2007 period. The CRSP file contains monthly stock prices, shares outstanding, trading volumes, and returns for all traded securities from 1926 to the present time.

There are a seemingly infinite number of financial variables that may be tested for statistical association with monthly security returns. Bloch et al. (1993) tested a set of fundamental variables for the US during the 1975–1990 period. Guerard (1997) tested a set of I/B/E/S variables for the 1982–1994 period. We will test the variables of these two studies using both fundamental and expectational data. We initially test the effectiveness of the individual variables using the information coefficients, ICs, rather than the upper quintile excess returns or the excess returns of individual variable portfolio optimizations. The information coefficient is the slope of the regression estimation in which ranked subsequent security returns are a function of

[5] Guerard (2006) re-estimated the GPRD model using PACAP data from the Wharton Research Data Services (WRDS) while teaching as an adjunct faculty member at the University of Pennsylvania. The WRDS/PACAP data are as close to the GPRD data as was possible in academia. The average, cross-sectional, quarterly WLRR model F-statistic in the GPRD analysis was 16 during the 1974–1990 period whereas the corresponding F-statistic reported in the Guerard (2006) was 11 for the post-publication, 1993–2001 period. Both sets of models were highly, statistically significant and could be effectively used as stock selection models. One finding of the Guerard updated analysis was the increasing role of analysts' revisions in Japan. In 1991, I/B/E/S covered 300 very large companies in Japan. By 2005, I/B/E/S covered over 3000 Japanese stocks, many of which had monthly and quarterly revisions that were more statistically, significantly associated with security total returns than was the case in 1991.

the ranked financial strategy. The advantage of the IC approach is that the slope has a corresponding t-statistic which allows one to test the null hypothesis that the strategy is uncorrelated with subsequent returns. In developing a composite model, one seeks to combine variables that are statistically associated with subsequent returns. The variables tested in this chapter are as follows:

EP = earnings per share/price per share;
BP = book per share/price per share;
CP = cash flow per share/price per share;
SP = sales per share/price per share;
DY = dividend yield, dividend per share/price per share;
NCAV = net current asset value, net current assets per share/price per share;
PM = price momentum, or $\text{price}_{t-1}/\text{price}_{t-12}$;
FEP1 = 1-year-ahead forecast earnings per share/price per share;
FEP2 = 2-year-ahead forecast earnings per share/price per share;
RV1 = 1-year-ahead forecast earnings per share monthly revision/price per share;
RV2 = 2-year-ahead forecast earnings per share monthly revision/price per share;
BR1 = 1-year-ahead forecast earnings per share monthly breadth/price per share; and BR2 = 2-year-ahead forecast earnings per share monthly breadth/price per share.

A consensus earnings-per-share I/B/E/S forecast, revisions and breadth variable, CTEF, was created and tested using I/B/E/S data since January 1976.

Table 3.1 shows the monthly ICs for all traded United States securities during the January 1985–December 2007 period. The majority of the variables are statistically associated with stockholder returns, a result consistent with the Bloch et al. (1993) and Guerard et al. (1997, 2009) studies. Moreover, Haugen and Baker (2009) employ many of these variables in their analysis contained in this volume. There is strong support for both fundamental variables (earnings, book value, cash flow,

Table 3.1 Monthly information coefficients, 1985–2007

Variable	IC (t)
EP	0.046 (53.65)
BP	0.011 (12.64)
CP	0.041 (47.34)
SP	0.009 (11.12)
DY	0.052 (40.57)
NCAV	−.015 (−17.09)
FEP1	0.036 (41.32)
FEP2	0.028 (29.82)
RV1	0.023 (25.20)
RV2	0.036 (41.32)
BR1	0.029 (31.71)
BR2	0.022 (23.40)
PM	0.044 (15.08)
CTEF	0.041 (45.21)

sales, dividends), earnings expectational variables in the anomalies literature (Levy 1999; Brown 2008; Ramnath et al. 2008), and intermediate price momentum variables (Brush 2001, 2007; Chan et al. 1996; Korajczyk and Sadka 2004; Fama and French 1992, 1995, 2008).[6]

The results of Table 3.1 support the estimation of the composite security valuation model reported in Guerard et al. (1997). That model incorporates reported earnings, book value, cash flow, and sales, the corresponding relative variables, and an equally weighted composite model of earnings forecasts, revisions, and breadth.

We estimate a similar monthly model for the January 1985–December 2007 period as follows:

$$\begin{aligned} \mathrm{TR}_{t+1} = {} & a_0 + a_1\mathrm{EP}_t + a_2\mathrm{BP}_t + a_3\mathrm{CP}_t + a_4\mathrm{SP}_t + a_5\mathrm{REP}_t + a_6\mathrm{RBP}_t \\ & + a_7\mathrm{RCP}_t + a_8\mathrm{RSP}_t + a_9\mathrm{CTEF}_t + a_{10}\mathrm{PM}_t + e_t, \end{aligned} \tag{3.3}$$

where

EP = [earnings per share]/[price per share] = earnings–price ratio;
BP = [book value per share]/[price per share] = book–price ratio;
CP = [cash flow per share]/[price per share] = cash flow–price ratio;
SP = [net sales per share]/[price per share] = sales–price ratio;
REP = [current EP ratio]/[average EP ratio over the past 5 years];
RBP = [current BP ratio]/[average BP ratio over the past 5 years];
RCP = [current CP ratio]/[average CP ratio over the past 5 years];
RSP = [current SP ratio]/[average SP ratio over the past 5 years];
CTEF = consensus earnings-per-share I/B/E/S forecast, revisions, and breadth,
PM = price momentum; and
e = randomly distributed error term.

The monthly ordinary least squares (OLS) regressions are plagued with approximately twice the number of observations outside the 95% confidence interval as one might expect given a normal distribution of residuals. These aberrant observations, or outliers, lead us to re-estimate the monthly regression lines using a Beaton–Tukey bi-weight (or robust, ROB) regression technique, in which each observation is weighted as the inverse function of its OLS residual. The application of the Beaton–Tukey ROB procedure addresses the issue of outliers. The weighted data is plagued with multicollinearity. Correlation among the independent variables may lead to statistically inefficient estimates of the regression coefficients.

[6] Fama and French have used a price momentum variable using the price 2 months ago divided by the price 12 months ago, thus avoiding the well-known return or residual reversal effect. The Brush, Korajczyk, and Fama studies find significant stock price anomalies, even with Korajcyk and Sadka using transactions costs. Lesmond et al. (2004) found that price momentum returns did not exceed transactions costs. The vast majority find that the use of 3, 6, and 12-month price momentum variables, often defined as intermediate term variables, are statistically (significantly) associated with excess returns. Brush (2001) reports that the quarterly IC of the 3-month price momentum variable exceeds the monthly IC, .073 vs.053. The fact that quarterly ICs often exceed monthly ICs was consistent with Bloch et al. (1993) analysis that fund that quarterly optimization and monthly rebalancing of portfolios maximized geometric means.

Bloch et al. (1993) and Guerard et al. (1993) applied latent-root regression (LRR) to the ROB-weighted data, referred to as weighted latent-root regression, WLRR, and produced models with higher in-sample F-statistics and higher out-of-sample geometric means. The average F-statistics in the WLRR regressions exceed the average OLS F-statistics by 35–40%, as was noted in Bloch et al. (1993). We create a composite model weight using the average weight of the positive coefficients of the preceding 12 monthly regressions, a monthly equivalent to the four-quarter averaging techniques used in Guerard et al. (1997). Of particular note is the large weighting of the CTEF variable, and the relatively small weighting of the BP variable, a result consistent with Guerard et al. (1997). In terms of information coefficients, ICs, the use of the WLRR procedure produces the highest IC for the models during the 1985–2007 period, shown in Table 3.2.

The WLRR technique produces the largest and most statistically significant IC; a result consistent with the previously noted studies and the GPRD example. The t-statistics on the composite model exceed the t-statistics of its components. The purpose of a composite security valuation model is to identify the determinants of security returns and produce a statistically significant, out-of-sample ranking metric of total returns.

An indication of the relative importance of the 10 individual variables in the composite is given by the regression coefficients in our 1985–2007 study. These time average values are illustrated in Charts 1–11. These results support the low P–E (high earnings yield) approach to value investing advocated by Graham et al. (1934, 1962) and validate as a cross-sectional return anomaly by Basu (1977). They also support the Fama and French (1992, 1995) finding that the book-to-market ratio is an important variable for explaining the cross-section of security returns. While EP and BP variables are significant in explaining returns, the majority of the forecast performance is attributable to other model variables, namely the relative earnings-to-price, the relative cash-to-price, relative sales-to-price, and earnings forecast variable. The second most statistically significant variable in the composite model is CTEF, shown in Chart 9 (Table 3.3).

Graphs of the monthly weights in the USER model are shown in Figs. 3.1–3.11. One can clearly see that variables go “in and out” of favor, as judged by the statically significant variables in the monthly cross-sectional regressions. However, the consensus earnings forecasting variable, CTEF, shown in Fig. 3.9 and the price momentum variable, PM, shown in Fig. 3.10, dominate the composite

Table 3.2 ICs of the composite security valuation model

Technique	IC (t)
Equal-weighted (EQ)	0.040 (47.35)
WLRR	0.047 (54.86)

Table 3.3 Time-average value of estimated coefficients

a1	a2	a3	a4	a5	a6	a7	a8	a9	a_{10}
.044	.038	.020	.038	.089	.086	.187	.122	.219	.224

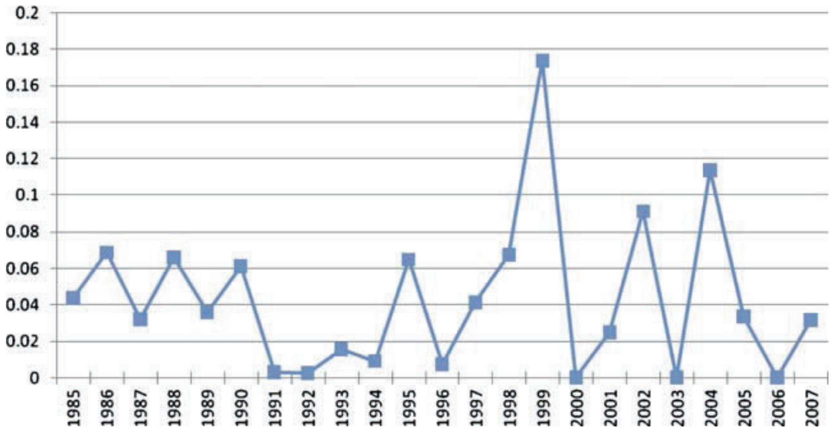


Fig. 3.1 USER Model, EP Weights

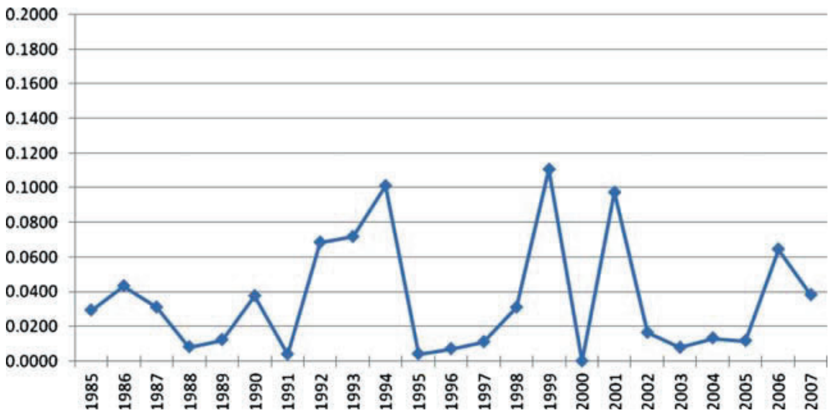


Fig. 3.2 USER Model, BP Weights

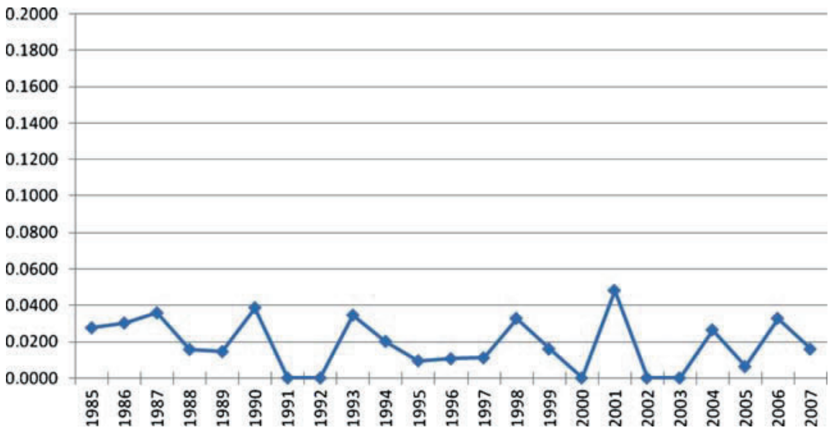


Fig. 3.3 USER Model, CP Weights

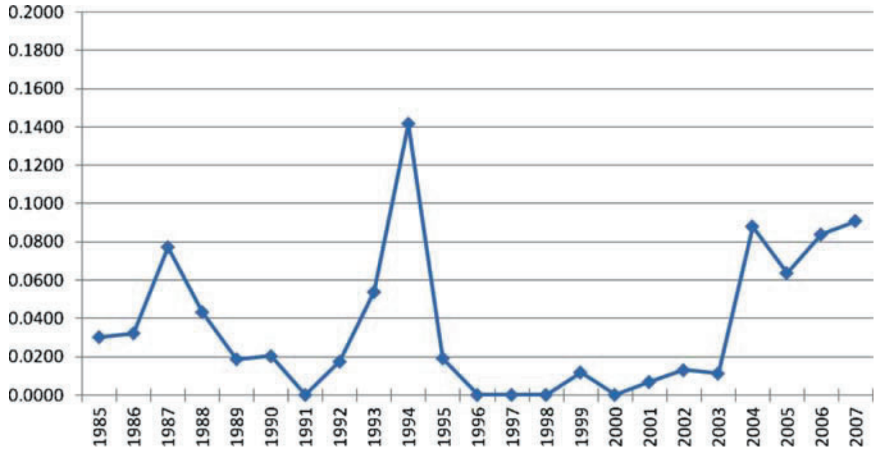


Fig. 3.4 USER Model, SP Weights

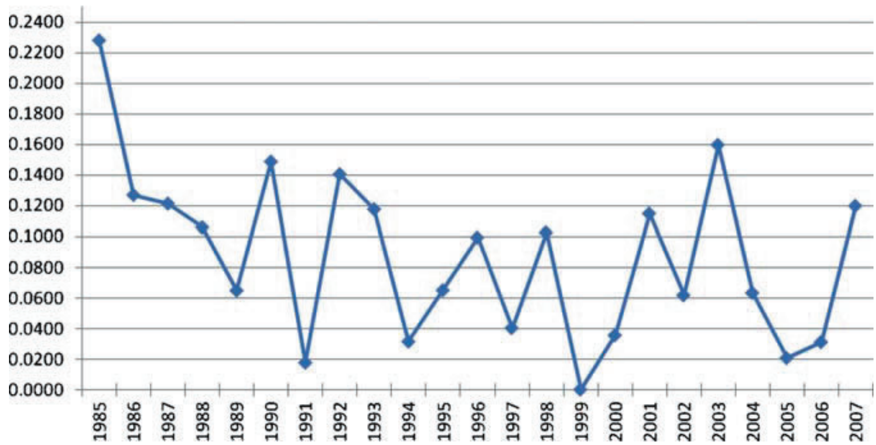


Fig. 3.5 USER Model, REP Weights

model, as is suggested by the fact that the variables account for 45% of the model average weights.

Employing the USER model as a "tilt factor", as in Guerard, Gultekin and Stone (1997) notation, an investment manager creates portfolios that look as close to the benchmark index as possible while maximizing the USER score. As risk-aversion level declines, the portfolio becomes more risky, the standard deviation rise, and there are fewer securities in the optimized portfolios. At the same time, returns should rise along with the Sharpe ratio and the information ratios, as compensation for bearing risk. Asset selection becomes more critical as fewer assets are selected for portfolios. Moreover, the market capitalization of securities decline in the BARRA-optimized portfolios and the size coefficient becomes more negative.

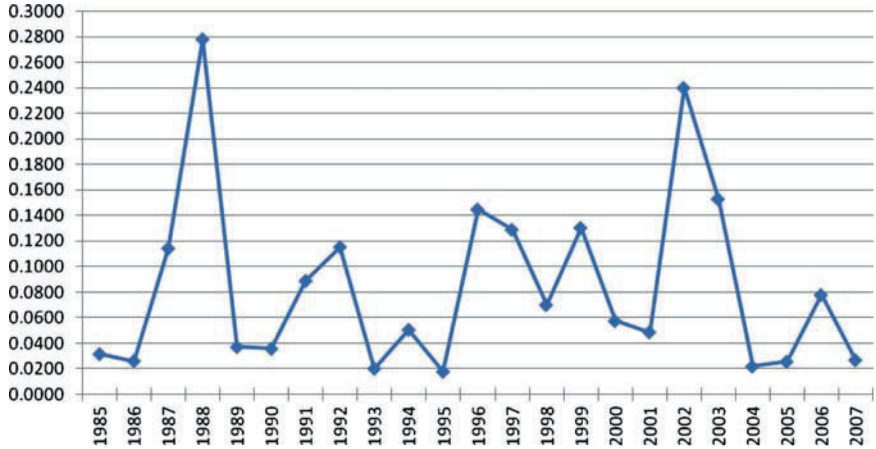


Fig. 3.6 USER Model, RBP Weights

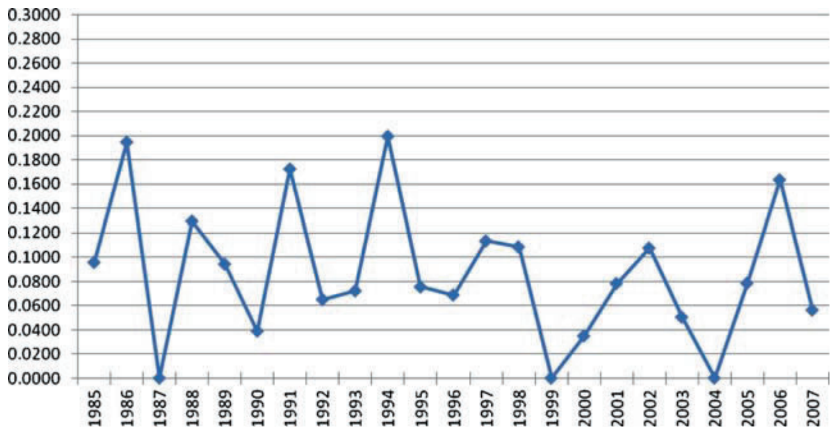


Fig. 3.7 USER Model, RCP Weights

The Sharpe ratio and the information ratio are maximized at a risk-aversion level of 0.02. Simulations are run from December 31, 1997 to November 30, 2007. Transactions costs of 125 basis points (1.25%) each way are taken from the portfolio returns, monthly reoptimization occur with an 8% turnover constraint.[7] While the maximum weight, the upper bound, is 4%, the lower bound is zero (no shorts). The USE BARRA efficient frontier is shown in Table 3.4 and illustrated graphically in Fig. 3.12.

[7] The USER analysis draws heavily from joint research being developed between Guerard et al. (2009). A working paper is under development.

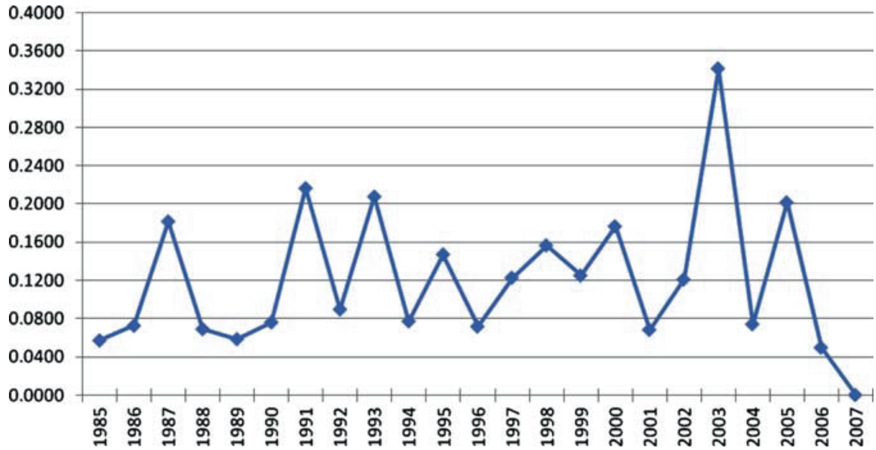


Fig. 3.8 USER Model, RSP Weights

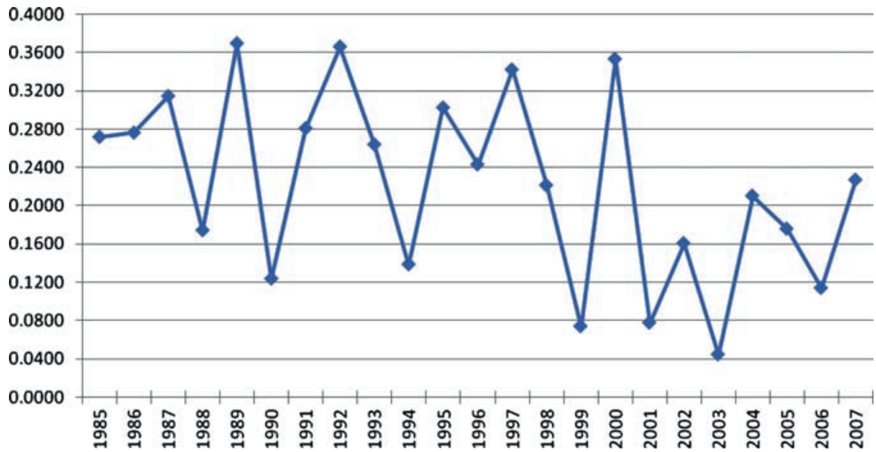


Fig. 3.9 USER Model, CTEF Weights

The USER model offers statistically significant ICs, excess returns, Sharpe ratios, and information ratios. Where are the sources of the USER model outperformance? The model offers active (above-market) exposures to smaller-capitalized (size) securities and securities in the nonestimation (Non-Est.) universe, earnings yield, volatility, momentum, and earnings variability. Only size and earnings yield exposures are rewarded with active returns that were statistically significant ($t > 1.645$, 10% level) (Table 3.5).

A global stock selection model can be estimated using the WorldScope database and the international I/B/E/S database for the January 1989–December 2008 period. The universe comprises all securities covered by at least one analyst. The

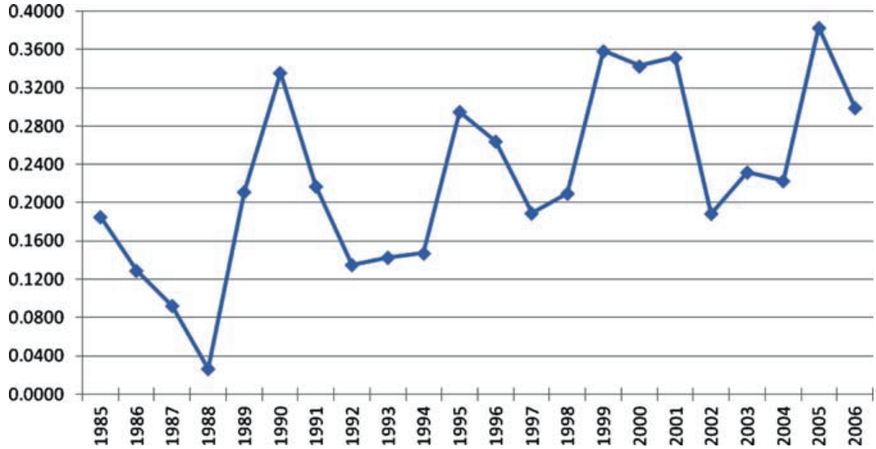


Fig. 3.10 USER Model, PM Weights

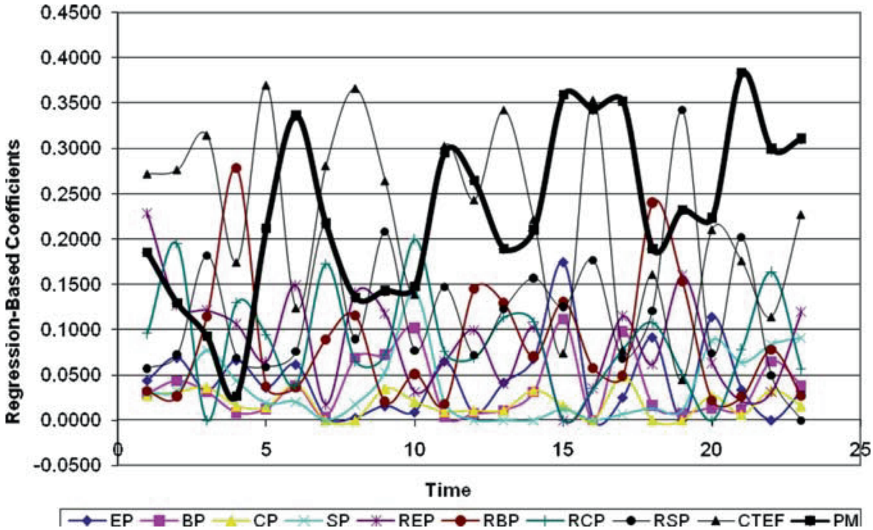


Fig. 3.11

information coefficients and composite variable weights are reasonably similar to the estimated US model. The information coefficients, ICs, are shown in Table 3.6.

The global stock selection average variable weights are shown in Table 3.7.

The average weight of the CTEF and PM-combined variables are approximately 45% in the US model, whereas the two variables are about 32.6% in the global model. The weight of the analysts' forecast variable constitutes the primary difference between the models (though the time periods are slightly different). The CTEF variable weight averages almost 22% in the US model whereas the CTEF is about

Table 3.4 Barra-estimated USER, EQ efficient frontiers, 1998.01–2007.12

RAL	Model	TManaged	STD	TActive	t-Active	AssetSel	t-AS	RskI	t-RskI	IR	Size
0.01	EQ	12.480	19.962	8.663	2.475	7.445	3.164	4.468	1.545	0.783	−1.176
0.01	USER	13.263	20.844	9.446	2.720	8.265	3.300	3.555	1.370	0.860	−1.220
0.02	EQ	10.993	20.193	7.176	2.267	6.526	2.980	3.759	1.458	0.717	−1.042
0.02	USER	12.517	20.494	8.700	2.779	7.490	3.281	3.214	1.394	0.879	−1.050
0.03	EQ	10.612	20.155	6.796	2.230	6.270	2.945	3.564	1.440	0.703	−0.990
0.03	USER	11.651	20.578	7.835	2.660	6.956	3.220	2.901	1.330	0.839	−0.980
0.05	EQ	9.155	19.992	5.339	2.054	5.300	2.848	2.640	1.280	0.650	−0.787
0.05	USER	9.355	20.346	5.539	2.360	5.742	3.226	2.014	1.234	0.745	−0.764
0.09	EQ	7.839	19.982	4.022	1.878	4.101	2.578	2.115	1.282	0.594	−0.582
0.09	USER	8.170	20.161	4.353	2.194	5.214	3.342	1.610	1.147	0.694	−0.598
0.15	EQ	6.540	19.979	2.723	1.575	3.265	2.408	1.629	1.233	0.498	−0.421
0.15	USER	7.122	19.800	3.305	1.967	4.358	3.197	1.349	1.169	0.622	−0.477
0.20	EQ	5.814	19.848	1.997	1.322	2.822	2.294	1.369	1.184	0.418	−0.356
0.20	USER	6.617	19.641	2.801	1.824	4.007	3.174	1.191	1.143	0.577	−0.416

where RAL = risk-aversion level, TManaged = total managed return, TActive = total active return, AssetSel = asset selection, IR = information ratio, Size, MOM, EY, value, and growth are BARRA multi-factor risk exposures, EQ = EQ-WT(EP, BP, CP, SP, REP, RBP, RCP, RSP, PM, CTEF), USER = WLRR-WT(EP, BP, CP, SP, REP, RBP, RCP, RSP, PM, CTEF)

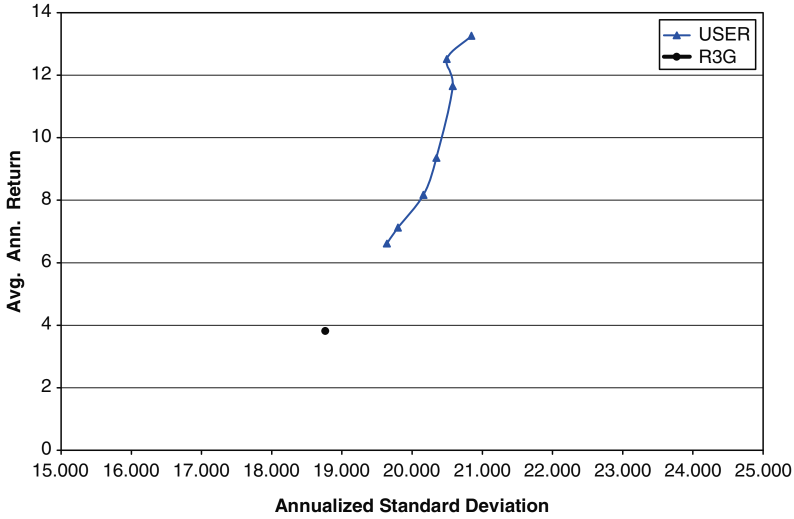


Fig. 3.12 USER model Efficient Frontier, 1/1998-12/2007

Table 3.5 Barra-optimized USER, 1/1998–12/2007

Source of return	Contribution	Risk	Info. ratio	T-statistic
Risk free	3.567	N/A	N/A	N/A
Benchmark	3.817	18.764		
Risk indices	3.214	6.521	0.440	1.394
Industries	0.752	3.511	0.200	0.633
Asset selection	7.490	7.020	1.037	3.281
Transactions costs	−2.756			
Total active	8.700	9.620	0.879	2.780
Total managed	12.517	20.494		
Contributions to risk index return				
	Avg. active exp.	Risk	Info. Ratio	T-statistic
Size	−1.050	4.211	0.703	2.224
Earnings yield	0.116	0.557	0.942	2.980
Volatility	0.253	2.109	0.272	0.860
Trading activity	−0.569	1.763	0.073	0.230
Currency sensitivity	0.052	0.376	0.031	0.099
Growth	0.044	0.329	−0.035	−0.111
Leverage	0.471	1.014	−0.040	−0.128
Value	0.541	1.079	−0.046	−0.146
Yield	0.209	0.513	−0.256	−0.810
Earnings variation	0.451	0.513	−0.168	−0.532
Momentum	0.290	2.283	−0.062	−0.196
Non-est. universe	0.438	3.379	−0.111	−0.353
Size nonlinearity	−0.399	1.877	−0.222	−0.703
Total				1.394

Benchmark is Russell 3000 growth index

Table 3.6 Global ICs, 1989.01–2008.12

Model	IC	T(IC)
EP	0.019	26.52
BP	−0.010	−11.43
CP	0.010	20.69
SP	−0.012	−14.18
DY	0.019	23.61
NCAV	−0.011	−16.88
FEP1	0.022	28.57
FEP2	0.020	26.54
RV1	0.017	22.22
RV2	0.021	27.51
BR1	0.026	28.00
BR2	0.027	28.21
PM	0.032	46.73
CTEF	0.030	39.95
EWC	0.023	27.70
WLRR	0.030	41.94

Table 3.7 Global stock selection variable weights, 1989–2008

Variable	Average weight
EP	0.083
BP	0.106
CP	0.060
SP	0.075
REP	0.048
RBP	0.056
RCP	0.039
RSP	0.105
CTEF	0.145
PM	0.181

14.5% in the global model. Price momentum and analysts' forecasts are the two most important variables, as they have the largest variable weights, in both models. Also remember that the US model and market capitalization constitutes approximately 40–50% of the global market capitalization.

3.3 Data Mining Corrections: Did You Get Lucky?

In the (practical) world of Wall Street, it is conventional wisdom to cut your historically backtested excess returns in half; that is, if your backtested excess return (the portfolio geometric mean return less the geometric mean of the benchmark) was 6%, or 600 basis points, an investment manager and/or a client might expect 3% excess returns in the future. In January 1991, Harry Markowitz and his Daiwa GPRD launched Fund Academy, a Japanese-only, Tokyo-based, investment strategy.

In its first year, ending December 1991, the Daiwa Portfolio Optimization System (DPOS) outperformed the benchmark by some 700 basis points. Markowitz asked "Did we get lucky?" The obvious answer is "No, we imported a US-estimated strategy run successfully at Drexel Burnham Lambert to Japan, its backtested excess returns were 50% higher in Japan than in the US, and its 700 basis points of excess returns were not much less than its 1,000 basis points of backtested outperformance during the 1974–1990 period." That answer, while quite true, was not scientific. Ganlin Xu, the Daiwa Global Research Department mathematician developed a testing methodology that Markowitz and Xu (1994) published in the *Journal of Portfolio Management*. Let us trace the development of the Markowitz and Xu model and estimate the Data Mining Corrections estimator for long-only and 130/30 strategies. In Bloch et al. (1993), some 200 historic US and Japanese equity model simulations were reported. The best backtested DPOS strategy was dominant in US and Japanese markets. The DPOS Japanese-only strategy was funded and the level of expected future outperformance was a question (always) asked by clients.

Let $\mathrm{GM_b}$ be the backtested geometric best of the "best" historical simulation during T periods. Markowitz and Xu (1994) work with the logarithm of the geometric mean as

$$g_{\mathrm{b}} = \log_e (1 + \mathrm{GM_b}) \tag{3.4}$$

The Markowitz and Xu data-mining corrections (DMC) test assumes that the T-period historical returns were identically and independently distributed (i.i.d.), and that future returns are drawn from the same population (also i.i.d.). As we test many models, not just the best model, the geometric mean is no longer the best unbiased estimate of the true, underlying population g_{b}.

Markowitz and Xu (1994) set y_{it} as the logarithm of one plus the return for the ith portfolio selection model in period t. y_{it} is of the form

$$y_t = \mu_i + z_t + \varepsilon_{it}, \tag{3.5}$$

where

μ_i is a model effect,
z_{t} is a period effect,
and $\varepsilon_{\mathrm{it}}$ is a random deviation.

Markowitz and Xu (1994) Model I assumes that z_{t} is observable and the return of a market index. In this case, r_{it} is an excess return of model I as

$$r_{it} = y_{it} - z_{it} = \mu_i + \varepsilon_{it}. \tag{3.6}$$

The random deviation $\varepsilon_{\mathrm{it}}$ has a zero mean and is uncorrelated with μ_{i} and other $\varepsilon_{\mathrm{js}}$ as

$$E(\varepsilon_{it}) = 0, \tag{3.7}$$

$$\operatorname{cov}(\mu_i, \varepsilon_{jt}) = 0 \quad \text{for all } j \text{ and } t, \tag{3.8}$$

and

$$\text{cov}(\varepsilon_{it}, \varepsilon_{js}) = 0 \quad \text{for all } i \neq j \text{ or } s \neq t. \tag{3.9}$$

Model I excess returns are assumed to be independent of time and the error terms of other models.

Markowitz and Xu (1994) Model II assumes that z_t, the market index return, is not observable. They did not estimate Model II in their original work, and we will not estimate it in this analysis. We proceed to estimate Model III. In Model III,

$$y_{it} = \mu_i + \varepsilon_{it}, \text{ as in (3.6).}$$

Markowitz and Xu did not require that models be independent of one another. Thus, in Model III, cov (ε_{it}, ε_{jt}) need not be zero. Thus, Model III is not only a general case (Model I being a special case of Model III), but is also consistent with testing in business-world portfolio construction and testing. Finally, the appropriate estimate of μ_{it} in Model I is not the average return

$$\bar{r}_i = \frac{\sum\limits_{t-1}^{T} y_{it}}{T}, \tag{3.10}$$

but rather

$$\bar{r} = \sum_{i=1}^{T} r_i/n. \tag{3.11}$$

The estimate of μ_i is regressed back to the average return (the grand average)

$$\hat{\mu} = \bar{r} + \beta\ (\bar{r}_i - \bar{r}), \tag{3.12}$$

where $0 < \beta < 1$.

The best linear estimate of the unknown μ_i, is

$$\hat{\mu}_i = E\mu + \beta\ (\bar{r}_i - E\mu) \tag{3.13}$$

and

$$\beta = \frac{\text{cov}(\bar{r}_i, \mu)}{\text{Var}(\bar{r}_i)}. \tag{3.14}$$

Thus, β is the regression coefficient of μ_i as a function of r_i. Does the return of the best model deviate (significantly) from the average model return? The best linear unbiased estimate of the expected model return vector μ is

$$\hat{\mu} = E(\mu)e + \text{Var}(\mu)\ \left[\tfrac{1}{T}C + \text{Var}(\mu)I\right]^{-1} \times\ (\tilde{y} - E(\mu)e)$$

and

$$C = \text{cov}(\varepsilon_e, \varepsilon_j).$$

The Markowitz–Xu DMC test did not use a "holdout period," as they can be routinely data mined as well. That is, one can vary the estimation and holdout periods to generate the desired conclusion. Markowitz and Xu (1994) tested the DPOS strategies in Bloch et al. (1993), and the best model is illustrated in Table 3.7. Table 3.7 is reprinted from Bloch et al. (1993). Markowitz and Xu reported a Model III β of 0.59, which was statistically significant; that is, approximately 59% of the excess returns could be expected to continue.

The MQ variable estimated in the Guerard et al. (2009) paper in this volume passes the Data Mining Corrections test criteria for both US and non-US markets, indicating that the stock selection and portfolio construction methodologies produce superior returns that are not due to chance. The MQ variable, when compared to the average of most models shown in Table 3.1, has a data-mining corrections coefficient of 0.47 and is highly statistically significant, having a F-value of 1.872. Thus, one could expect 47% of the excess returns of the MQ model relative to the average return to be continued. Backtesting can never be perfect but it can be statistically significant.

The reader notes that in the Markowitz GPRD analysis, sophisticated regression techniques produce higher geometric means than the (simple) OLS regression technique.

Table 3.8 D-POS, Japan

SID	OP	TOV	RP	Rtri	ERET	GM	Shrp
900869	3	10.00	1	25.0	PROPRIETARY	25.60	0.89
900867	3	10.00	0	0.0	PROPRIETARY	25.39	0.89
900868	3	10.00	1	20.0	PROPRIETARY	25.32	0.87
900860	3	15.00	1	25.0	PROPRIETARY	24.28	0.84
900853	3	15.00	0	0.0	PROPRIETARY	24.19	0.85
900858	3	10.00	1	25.0	PROPRIETARY	24.04	0.85
900852	3	12.50	0	0.0	PROPRIETARY	23.94	0.85
900855	3	10.00	1	20.0	PROPRIETARY	23.93	0.86
900856	3	12.50	1	20.0	PROPRIETARY	23.90	0.84
900854	3	17.50	0	0.0	PROPRIETARY	23.89	0.84
900859	3	12.50	1	25.0	PROPRIETARY	23.89	0.83
900857	3	15.00	1	20.0	PROPRIETARY	23.81	0.82
900819	3	10.00	0	0.0	REGR(WLRR,4Q,4)	22.74	0.83
900820	3	10.00	1	25.0	REGR(WLRR,4Q,4)	22.68	0.82
900944	3	10.00	0	0.0	BPR	22.43	0.78
900908	3	10.00	1	20.0	REGR(LRR,4Q,9.1)	22.23	0.75
900874	3	10.00	0	0.0	REGR(OLS,4Q,8)	22.16	0.79
900878	3	10.00	0	0.0	REGR(OLS,4Q,9.1)	22.16	0.79
900903	3	10.00	0	0.0	REGR(OLS,4Q,8)	22.16	0.79
900914	3	10.00	0	0.0	REGR(OLS,4Q,9.1)	22.16	0.79
900841	3	10.00	1	25.0	REGR(WLRR,1Q,4)	22.00	0.79
900817	3	10.00	0	0.0	REGR(LRR,4Q14)	21.99	0.76

(continued)

Table 3.8 (continued)

SID	OP	TOV	RP	Rtri	ERET	GM	Shrp
900983	3	10.00	1	20.0	REGR(WLRR,4Q,9.1)	21.93	0.75
900984	3	10.00	1	20.0	REGR(WLRR,4Q,9.1)	21.86.	0.75
900794	3	15.00	1	20.0	REGR(WLRR,1Q,4)	21.84	0.76
900818	3	10.00	1	25.0	REGR(LRR,4Q,4)	21.84	0.75
900877	3	10.00	0	0.0	REGR(WLRR,4Q,8)	21.84	0.78
900906	3	10.00	0	0.0	REGR(WLRR,4Q,8)	21.84	0.78
900985	3	12.50	1	20.0	REGR(WLRR,4Q,9.1)	21.84	0.75
900913	3	10.00	0	0.0	REGR(WLRR,4Q,9.2)	21.83	0.77
900793	3	12.50	1	20.0	REGR(WLRR,1Q,4)	21.78	0.78
900791	3	12.50	0	0.0	REGR(WLRR,1Q,4)	21.75	0.79
900792	3	15.00	0	0.0	REGR(WLRR,1Q,4)	21.68	0.77
900982	3	10.00	1	20.0	REGR(WLRR,4Q,9.1)	21.66	0.75
900842	3	10.00	1	25.0	REGR(WLRR,10,4)	21.55	0.79
900766	3	10.00	1	20.0	REGR(WLRR,1Q,4)	21.49	0.78
900810	3	15.00	0	0.0	REGR(WLRR,1Q,4)	21.47	0.76
900901	3	10.00	0	0.0	REGR(LRR,4Q,9.1)	21.45	0.72
900813	3	10.00	0	0.0	REGR(OLS,4Q,4)	21.42	0.78
900840″	3	10.00	1	25.0	REGR(WLRR,1Q,4)	21.41	0.76
900838	3	10.00	1	25.0	REGR(WLRR,1Q,4)	21.40	0.76
900909	3	10.00	1	20.0	REGR(WLRR,4Q,9.1)	21.40	0.75
900910	3	10.00	0	0.0	REGR(LRR,4Q,9.2)	21.34	0.75
900816	3	10.00	1	25.0	REGR(ROB,4Q,4)	21.30	0.76
900839	3	10.00	1	25.0	REGR(WLRR,1Q,4)	21.30	0.75
900912	3	10.00	0	0.0	REGR(LRR,4Q,9.2)	21.29	0.71
900765	3	10.00	0	0.0	REGR(WLRR,1Q,4)	21.24	0.76
900815	3	10.00	0	0.0	REGR(ROB,4Q,4)	21.23	0.76
900902	3	10.00	0	0.0	REGR(WLRR,4Q,9.1)	21.16	0.74
900986	3	15.00	1	20.0	REGR(WLRR,4Q,9.1)	21.09	0.72
900954	3	10.00	0	0.0	REGR(OLS,4Q,4)	20.91	0.72
900876	3	10.00	0	0.0	REGR(LRR,4Q,8)	20.90	0.74
900905	3	10.00	0	0.0	REGR(LRR,4Q,8)	20.90	0.74
900911	3	10.00	0	0.0	REGR(ROB,4Q,9.2)	20.66	0.72
900907	3	10.00	1	20.0	REGR(ROB,4Q,9.1)	20.36	0.74
900763	3	10.00	0	0.0	REGR(LRR,1Q,4)	20.21	0.71
900875	3	10.00	0	0.0	REGR(ROB,4Q,8)	20.15	0.71
900904	3	10.00	0	0.0	REGR(ROB,4Q,8)	20.15	0.71
900787	3	12.50	0	0.0	REGR(LRR,1Q,4)	20.08	0.71
900900	3	10.00	0	0.0	REGR(ROB,4Q,9.1)	20.07	0.72
900781	3	12.50	1	20.0	REGR(OLS,1Q,4)	19.96	0.71
900788	3	15.00	0	0.0	REGR(LRR,1Q,4)	19.92	0.70
900764	3	10.00	1	20.0	REGR(LRR,1Q,4)	19.88	0.70
900790	3	15.00	1	20.0	REGR(LRR,1Q,4)	19.81	0.70
900789	3	12.50	1	20.0	REGR(LRR,1Q,4)	19.78	0.70
900779	3	12.50	0	0.0	REGR(OLS,1Q,4)	19.77	0.67
900786	3	15.00	1	20.0	REGR(ROB,1Q,4)	19.76	0.71
900780	3	15.00	0	0.0	REGR(OLS,1Q,4)	19.72	0.69

(continued)

Table 3.8 (continued)

SID	OP	TOV	RP	Rtri	ERET	GM	Shrp
900784	3	15.00	0	0.0	REGR(ROB,1Q,4)	19.67	0.71
900782	3	15.00	1	20.0	REGR(OLS,1Q,4)	19.41	0.69
900759	3	10.00	0	0.0	REGR(OLS,1Q,4)	19.40	0.67
900785	3	12.50	1	20.0	REGR(ROB,1Q,4)	19.33	0.69
900760	3	10.00	1	20.0	REGR(OLS,1Q,4)	19.31	0.66
900783	3	12.50	0	0.0	REGR(ROB,1Q,4)	19.10	0.69
900761	3	10.00	0	0.0	REGR(ROB,1Q,4)	19.03	0.68
900931	3	10.00	0	0.0	CPR	19.01	0.68
900762	3	10.00	1	20.0	REGR(R0B,1Q,4)	19.00	0.67
900932	3	10.00	0	0.0	SPR	18.63	0.61
900716	3	10.00	1	20.0	Benchmark	17.25	0.60
900927	3	10.00	0	0.0	EPR	16.82	0.57
900826	6	20.00	3	25.0	PROPRIETARY	24.63	0.84
900709	6	20.00	3	20.0	PROPRIETARY	23.61	0.81
900710	6	20.00	3	25.0	PROPRIETARY	23.44	0.82
900733	6	25.00	1	20.0	PROPRIETARY	23.34	0.80
900773	6	17.50	3	20.0	PROPRIETARY	23.26	0.78
900707	6	20.00	3	20.0	PROPRIETARY	23.08	0.79
900847	6	20.00	0	0.0	PROPRIETARY	22.62	0.81
901030	6	20.00	3	20.0	BPR	22.42	0.78
900796	6	20.00	3	20.0	REGR(OLS,2S,4)	22.33	0.79
901047	6	20.00	3	20.0	BPR	22.20	0.77
900770	6	22.50	0	0.0	REGR(OLS,1S,4)	22.17	0.77
900795	6	20.00	0	0.0	REGR(OLS,2S,4)	22.14	0.79
900749	6	20.00	3	25.0	REGR(OLS,1S,4)	22.03	0.78
900800	6	20.00	3	0.0	REGR(LRR,2S,4)	21.98	0.78
900849	6	20.00	0	0.0	REGR(LRR,3S,4)	21.98	0.77
900748	6	20.00	3	20.0	REGR(OLS,IS,4)	21.80	0.77
900754	6	20.00	3	20.0	REGR(LRR,12,4)	21.68	0.74
900747	6	20.00	0	0.0	REGR(OLS,1S,4)	21.65	0.77
900802	6	20.00	3	20.0	REGR(WLRR,2S,4)	21.60	0.79
901029	6	20.00	0	0.0	BPR	21.59	0.76
900755	6	20.00	3	25.0	REGR(LRR,1S,4)	21.52	0.74
900799	6	20.00	0	0.0	REGR(LRR,2S,4)	21.51	0.77
901046	6	20.00	0	0.0	BPR	21.49	0.76
900801	6	20.00	0	0.0	REGR(WLRR,2S,4)	21.40	0.79
900769	6	17.50	0	0.0	REGR(OLS,1S,4)	21.34	0.76
900778	6	22.50	3	20.0	REGR(WLRR,1S,4)	21.30	0.79
900772	6	22.50	0	0.0	REGR(LRR,1S,4)	21.26	0.72
900753	6	20.00	0	0.0	REGR(LRR,lS,4)	21.20	0.72
900756	6	20.00	0	0.0	REGR(WLRR,1S,4)	21.10	0.77
900757	6	20.00	3	20.0	REGR(WLRR,1S,4)	21.06	0.78
900758	6	20.00	3	25.0	REGR(WLRR,1S,4)	21.02	0.78
900777	6	17.50	3	20.0	REGR(WLRR,1S,4)	20.95	0.78
900771	6	17.50	0	0.0	REGR(LRR,1S,4)	20.93	0.71
900848	6	20.00	0	0.0	REGR(ROB,3S,4)	20.74	0.76

(continued)

Table 3.8 (continued)

SID	OP	TOV	RP	Rtri	ERET	GM	Shrp
900776	6	22.50	3	20.0	REGR(ROB,1S,4)	20.37	0.76
900797	6	20.00	0	0.0	REGR(ROB,22,4)	20.24	0.75
900798	6	20.00	3	20.0	REGR(ROB,2S,4)	20.12	0.76
900752	6	20.00	3	25.0	REGR(ROB,1S,4)	19.56	0.73
900751	6	20.00	3	20.0	REGR(ROB,1S,4)	19.35	0.73
900750	6	20.00	0	0.0	REGR(ROB,lS,4)	19.29	0.72
900775	6	17.50	3	20.0	REGR(ROB,1S,4)	19.26	0.72
901049	6	20.00	3	20.0	CPR	18.99	0.68
901051	6	20.00	3	20.0	SPR	18.69	0.61
901048	6	20.00	0	0.0	CPR	18.65	0.67
901050	6	20.00	0	0.0	SPR	17.87	0.59
901045	6	20.00	3	20.0	EPR	17.55	0.59
901044	6	20.00	0	0.0	EPR	17.34	0.59

Source: Bloch et al. (1993).
Simulation results: sorted by geometric mean. Let UL = 2.0, TCR = 2.0, PM = −1, PPar = 0.90, Begin = 7412, and End = 9012. SID = simulation ID; OP = period of reoptimization; TOV = turnover constraint; Rtri = rebalancing trigger; ERET = model description; GM = geometric mean; and Shrp = Sharpe ratio. *REGR (technique, period, equation)*. Technique = OLS for ordinary least-squares regression analysis, LRR for latent-root regression, ROB for robust regression, WLRR for weighted latent-root regression; period = 1S for one-period semiannual analysis, 2S for two-period semiannual analysis, 3S for three-period semiannual analysis, 1Q for one-period quarterly analysis, 4Q for four-period quarterly analysis; equation = 4; TRR $= a_0 + a_1$EPR $+ a_2$BPR $+ a_3$CPR $+ a_4$SPR $+ a_5$REPR $+ a_6$RPBR $+ a_7$RCPR $+ a_8$RSPR $+ e_t$. 8; TRR $= a_0 + a_1$EPR $+ a_2$BPR $+ a_3$CPR $+ a_4$SPR $+ e_t$. 9.1; TRR $a_0 + a_1$EPR $+ a_2$BPR $+ a_3$CPR $+ a_4$SPR $+ a_5$REPR(2) $+ a_6$RPBR(2) $+ a_7$RCPR(2) $+ a_8$RSPR(2) $+ e_t$, where (2) denotes 2-year averages of relative variables. 9.2; TRR $= a_0 + a_1$EPR $+ a_2$BPR $+ a_3$CPR $+ a_4$SPR $+ a_5$REPR(3) $+ a_6$RPBR(3) $+ a_7$RCPR(3) $+ a_8$RSPR(3) $+ e_t$, where (3) denotes 3-year averages of relative variables

3.4 The JLM Market Simulator

In the early 2000s, Harry worked in joint research with Bruce Jacobs and Ken Levy of Jacobs Levy Equity Management, a provider of quantitative equity strategies for institutional clients, where he helped to construct the JLM Market Simulator (Jacobs et al. 2004). The JLM Sim is an asynchronous simulation that investors can use to create a model of the market using their own inputs. The investor's portfolio selection choice comes from the risk-aversion coefficient parameter that helps the client choose from a desired portfolio on the efficient frontier. The JLM Simulator was described in an article in the 30th anniversary issue of *The Journal of Portfolio Management*. The JLM Sim makes extensive use of Markowitz's experiences in programming and simulation from his years working on SIMSCRIPT, a simulation-programming language that Markowitz developed at RAND in the early 1960s and later formed the basis of CACI Products Company (Markowitz 2002; Markowitz et al. 1963).

JLM Sim employs an entity, attribute, and set structure. For example, one entity is the simulation system as a whole. Attributes of the system include the current lending and borrowing rates of interest. Sets owned by the system include all the securities, statisticians, and portfolio analysts in the simulation, as well as the number of kept trading days, the number of kept months, the investor templates, and the trader templates. Each simulated investor is created from an investor template with attributes including a risk-aversion parameter (much like the GPRD "pick parameter" or Barra risk-aversion level previously discussed), and reoptimization frequency. At the time of reoptimization, for each investor template, portfolio analysts choose the desired efficient portfolios from the mean-variance efficient frontier. For constructing the efficient frontiers, portfolio analysts obtain their estimates of expected security returns and covariances from statisticians. Traders attempt to execute trades to achieve the investor's desired portfolio. Every trader is created from a trader template, whose attributes determine behavior such as how aggressively securities should be traded. Securities are entities with attributes including the last traded price, the current price, the price at the start of the day, the daily volume, and the price at the start of the month.

In JLM Sim, events change the state, or status, of the simulated market. Events cause future event occurrences. There are four events in JLM Sim. In the initialize event, or initialization, investor templates, trader templates, statisticians, and portfolio analyst templates are created with attributes specified by the user. Statisticians may use specified, expected return procedures or historical returns to estimate expected returns, and historical data are used to estimate the covariance structure. The investor starting wealth is determined during initialization using random draws from a lognormal distribution of user-specified parameters (attributes of the investor template). Presimulation security returns are randomly generated from a user-specified factor model. The investor template includes the reoptimization period as an attribute.

In the second event, reoptimization, portfolio analysts use statistical estimates to generate ideal portfolios on the efficient frontier for each investor template. Orders are placed with traders, which could be matched by other-sided orders with any balance placed on the books for execution. The trader initially places a buy order with a limit price, bidding less than the current price. The price could be the average of the bid and the ask prices. A transaction occurs if the sell-order set contains a sell price at or less than the buyer's limit price. The transaction quantity is the lesser of the seller's and buyer's desired trade sizes. Buy transactions not completed are entered into buy-order sets.

The third event is the review order event, in which orders are either repriced or canceled. Repricing leads JLM Sim to consider matching orders. If the order is not completed after a consideration of matching orders, it is placed on the books and a new order review is scheduled. The fourth event, the end-of-day event, updates daily (and other frequency, i.e., monthly, quarterly) statistics. Accounts are marked to market with leverage, and accounts violating maintenance margin requirements are turned over to a liquidation trader, which is a trader that performs quick, not necessarily favorable, execution.

The JLM Sim runs on a standard personal computer using the Windows operating system. It uses random access memory (RAM) rather than disk memory, which speeds up computation time. Moreover, it uses specialized (EASE-E) software to file, remove, and find large ordered sets that are stored in RAM.

3.5 Overview and Estimation of the 130/30 Strategy

Most investment managers only purchase stocks and seek to outperform the market. They are referred to as long-only managers. Managers who borrow stocks from a broker to sell immediately and repurchase later are called short sellers. Short sellers benefit when stock prices fall. A relatively, recently developed strategy uses 130% of account equity to buy stocks and 30% of account equity to short sell in order to finance the additional stock purchase. Such a portfolio has a net exposure of 100% and is known as a 130/30 portfolio.

The initial impetus for 130/30 portfolios came from Jacobs et al. (1998), which extended the concept of short-selling beyond market-neutral portfolios to include long–short portfolios that maintain a full-market exposure. Long–short portfolios with any given exposure to the underlying market benchmark should be constructed with an integrated optimization that considers simultaneously both long and short positions and the benchmark asset. Rather than combining a long-only portfolio with a market-neutral portfolio, it is better to blend active long and short positions so as to obtain a desired benchmark exposure. Jacobs, Levy, and Starer laid the foundation for optimally equitizing an active, long–short portfolio when exposure to a benchmark is desired. Jacobs and Levy (2005b) highlighted the advantages of 130/30 portfolios over long-only and other long–short approaches, and also some of the differences between 130/30 and long-only and market-neutral, long–short portfolios.

A 130/30 portfolio simultaneously secures both long and short exposures, and relaxes the long-only constraint required in more traditional investing.[8] This is how the strategy typically works: a 130/30 portfolio invests 130% of the account equity in long stock positions and 30% in short stock positions. The 30%-leveraged long positions are financed from the proceeds of the short sales. For example, a $100-million account would hold $130 million of long stocks. The $30-million leveraged portion would be financed through $30 million of short-sale proceeds.

The net exposure of the portfolio is 100%, or $100 million, as shown in Fig. 3.13. As such, a 130/30 portfolio is expected to generate returns of the same order of magnitude as the underlying universal benchmark. Because of this expected return structure, and because the net exposure remains 100%, a 130/30 portfolio is

[8] One who buys a stock, or is long a stock, expects the stock's price to rise. Conversely, one who short sells a stock wants that stock's price to fall. For example, a manager short sells a stock by borrowing that stock from a broker, selling it at say, $100; then replaces the broker's stock by repurchasing it later at a lower price, say $75. As a short seller, that manager earns the difference of $25 ($100 minus $75). However, the manager loses money if the shortened stock rises instead of falls in value.

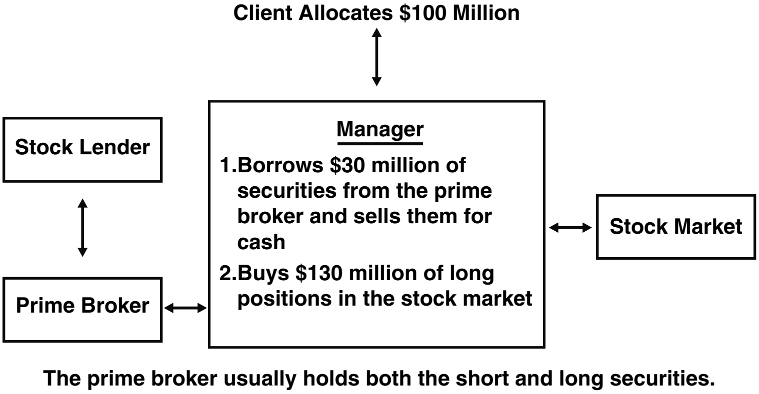


Fig. 3.13 Funding a typical 130/30 portfolio

viewed as a substitute for a long-only portfolio allocation. According to Park and Bucknor's (2006) survey results, a 130/30-style, long–short portfolio is positioned by plan sponsors in the normal investment categories rather than in the alternative investment, or absolute return, bucket.

Similar portfolios with similar characteristics and definitions can be formed as 110/10, 120/20, or 140/40 portfolios. Perhaps because of regulation T, which caps the amount of credit brokers, and therefore, dealers are allowed to extend to clients, a 150/50 formulation seems to be the limit before investors shift the allocation into the alternatives class. Regulation T establishes the margin on stock, and the Federal Reserve has determined certain margin requirements.[9] The underlying universe of stocks and related benchmark can be any traditional equity bucket incorporating large/small, growth/value, international/domestic, or developed/emerging, market allocation considerations.

A 130/30 portfolio is an active strategy that can be used by a skilled manager to increase the expected portfolio return. For example, suppose that a particular equity manager with skill in a certain universe has been able to generate excess returns of 4% on average each year. Now, consider the expected results using the 130/30 format. The manager is expected to generate excess return of 4% on an unleveraged long portfolio; it would be reasonable to assume an excess return of 5.2% on a portfolio with 1.3:1 long-side leverage. If the manager has an equal skill on the short side, an additional excess return of 1.2% could be expected on the 30% short positions. Accordingly, the expected excess returns for the 130/30 portfolio would be 6.4% (4.0×1.6).

Moreover, a manager who thinks a smaller capitalization stock will underperform can only slightly underweight the security in a long-only portfolio as the underweight is constrained by the stock's weight in the benchmark. But the manager who

[9] Please see Jacobs et al. (2006) for a discussion of the regulation tissues in long–short portfolios.

can sell short in a 130/30 portfolio can attain a larger underweight.[10] What is important is that the manager be able to identify stocks that will underperform the market (universal benchmark). Thus, a manager with stock-selection skill could earn higher returns at the same magnitude of market-based or beta risk. Investors should note that market risk is not the sole source of risk; indeed, there are many other sources of risk to consider.

In the example just described, excess return is assumed to be linearly related to exposure, though this might not actually be the case. It would seem that a linear relationship on the long side could be achieved simply by increasing the level of exposure to the same stocks. For example, if a long-only portfolio contained 100,000 shares of a certain stock, then a 130/30 variant would have 130,000 shares of that stock. However, the manager may have already reached a capacity in the security. If so, the manager might add a new stock rather than increase the weight in the current stock. Such a strategy might also be adopted to control risk. Depending on the manager's forecasting model, adding new securities might require moving into lower-ranked, less-favored stocks. This effect likely increases as the format is changed from 110/10 to 130/30 to 150/50 and more funds become available for investing long. So, instead of the long-side excess return increasing linearly, the move to a long–short strategy might result in a diminishing pattern of returns, though possibly with a more favorable risk profile.

A long-only portfolio does not allow a skillful manager to fully exploit negative opinions to generate excess returns because the most extreme underweight position the manager can take is to not hold a stock at all (0% weight). With a 130/30 portfolio, however, a manager has more leeway to produce excess returns from negative opinions. Suppose a stock comprises a weight of only 10 basis points in the benchmark. The most a long-only manager with a negative view on the stock can underweight that stock is by 10 basis points, through avoiding its purchase. However, that manager can capture a higher return on a positive view because the same stock held at 2% is a 190-basis-point overweight. As shorting gives a manager the ability to act on a significantly negative view, shorting introduces more flexibility and potentially provides a mechanism for important gains. Jacobs, Levy, and Markowitz, which we denote as JLM (2005a, b, 2006), discuss the relative advantage of shorting smaller capitalized stocks at a great length.

JLM (2005a, b, 2006) present an interesting "$2n$" representation for their long–short portfolios. We adopt the JLM notation and compare their approach with the Guerard et al. (2009) 130/30 portfolio construction analysis in the volume.

A portfolio is constructed with n securities having expected returns $\mu_1, \mu_2, \ldots, \mu_n$. The reader recalls from Chaps. 1 and 2 that the portfolio expected return, E_p, is a weighted sum of the n security returns and their respective weights, $x_1, x_2, \ldots, x_n$ is

$$E_\mathrm{p} = \sum_{i=1}^{n} x_i \mu_i . \tag{3.15}$$

[10] Clarke et al. (2004).

The portfolio variance can be written as

$$V_{\mathrm{p}} = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j \sigma_{ij}. \tag{3.16}$$

A long-only portfolio may use several (m) constraints as

$$\sum\nolimits_{k=1}^{n} a_{jk} x_k = b_j \quad \text{for } j = 1, \ldots, m, \tag{3.17}$$

and

$$x_i \geq 0. \tag{3.18}$$

Furthermore,

$$\sum_{i=1}^{n} x_i = 1. \tag{3.19}$$

Now, the investment is fully invested.

A long–short portfolio does not impose a constraint (3.18). That is, x_i need not be positive. A negative x_i is a short. Long–short portfolios, as we see in Guerard et al. (2009), contain n variables long and n securities short in the same set of securities as

$$\sum\nolimits_{k=1}^{2n} a_{jk} x_k = b_j \quad \text{for } j = 1, \ldots, m \tag{3.20}$$

and

$$x_i \geq 0, \quad \text{for } i = 1, \ldots, 2n. \tag{3.21}$$

JLM assume long-only portfolios are subject to constraints (3.17) and (3.18) and long–short portfolios are subject to (3.20) and (3.21). We have discussed how upper and lower bounds are present in (3.17) and (3.18). An upper bound, μ, on short-selling security i, as used in Guerard et al. (2009), is imposed by setting $b_j = \mu$ and $a_{jk} = 1$. Regulation T of the United States Federal Reserve, as discussed before, requires that the sum of the long and short (absolute values of the shorts) does not exceed twice the account equity as

$$\sum\nolimits_{i=1}^{2n} x_i \leq H. \tag{3.22}$$

x_i represents long positions for longs, L, for $i \in [1, n]$ and shorts, S, for $i \in [n+1, 2n]$. H is now 2, or a 50% margin. Jacobs et al. (2006) discuss how an enhanced, active equity (e.g., 130/30) strategy has an exposure ν such that

$$\left| \left(\sum\nolimits_{i \in L} x_i - \sum\nolimits_{1 \in S} x_i \right) - \upsilon \right| \leq \tau, \tag{3.23}$$

for small, non-negative tolerance level denoted by τ. Full market exposure is achieved using $\nu = 1$.

Let r_i be the return for security i and r_c be the return on cash (or collateral). The return of the long–short portfolio return may be written as

$$E_p = \sum_{i=1}^{2n} x_i \mu_i \sum_{i=n+1}^{2n} (-r_{i-n}) x_i + r_c \sum_{i=n+1}^{2n} h_{i-n} x_i. \tag{3.24}$$

JLM remind the reader that the third term is the short rebate as

$$E_p = \sum_{i=1}^{2n} x_i \mu_i. \tag{3.25}$$

Equation (3.25) is the expected return of the long–short portfolio. In a multifactor (APT) model, where

$$r_i = \alpha_i + \sum_{k=1}^{K} \beta_{ik} f_k + u_i, \tag{3.26}$$

K is the number of factors with factor loadings β_{ik} and u_i is idiosyncratic risk, as discussed in Chap. 2.

Portfolio optimization generally requires the inversion of a covariance matrix of security returns. Securities with returns having nonzero covariances lead to "dense" covariance matrices, which are computationally demanding to invert. In contrast, as shown in Chap. 1, inversion of "nearly diagonal matrices" can be performed with little effort. Jacobs et al. (2005) show how modeling techniques can be used to convert a dense covariance matrix into a nearly diagonal one. To do this, they create s fictitious securities with weights $y_1, \ldots, y_s$ defined as follows:

$$y_s = \sum_{i=1}^{n} (\mu_{is} - \mu_i) x_i, \text{ for } s = 1, \ldots, s. \tag{3.27}$$

The portfolio variance can then be written as

$$V_p = \sum_{i=1}^{n} x_i^2 V_i + \sum_{s=1}^{s} y_s^2 \, p_s, \tag{3.28}$$

where

$$V_i = \sum_{s=1}^{s} p_s V_{is}.$$

A long-only portfolio variance may be written as

$$V_p = \sum_{i=1}^{n} x_i^2 V_i + \sum_{s=1}^{s} y_k^2 \, W_k, \tag{3.29}$$

and the long–short portfolio variance may be written as

$$V_p = \sum_{i=1}^{2n} x_i^2 V_i + \sum_{s=1}^{s} y_k^2 W_k - 2 \sum_{i=1}^{n} x_i x_{n+1} V_i. \tag{3.30}$$

JLM define a "trim" portfolio as having no longs and shorts in the same security (simultaneously).

JLM show that standard, long-only portfolio optimizers can be used for long–short portfolios using their $2n$ representation, as long as the covariance model satisfies a "trimability" condition. An efficient portfolio can be "trimmed" if simultaneous long and short positions in the same security can be eliminated while maintaining feasibility and not reducing a portfolio expected return. Simultaneous long and short positions in a security are rarely possible in Guerard et al. (2009) because of the portfolio tilt variable.

The 130/30 strategy affords the manager the opportunity to avoid much of the loss of portfolio construction efficiency. One way to measure this efficiency is by using the transfer coefficient, which is the correlation between active weights, and forecast residual returns in portfolio construction. The closer the transfer coefficient comes to 1.0, the more efficient is the portfolio construction. Clarke et al. (2002) show that portfolio constraints, such as using long-only strategies, lower the transfer coefficient from 0.98 to 0.58, with a 5.0 tracking error.

Within broad parameters, a manager with forecasting skill should be able to use a 130/30 format to increase the expected returns compared with a long-only offering. McKinley Capital tested a simulated 130/30 portfolio against its Non-US-developed growth (long-only) Portfolio; the simulated portfolio had a higher Sharpe ratio than the long-only portfolio.[11] Thus, the return-to-risk ratio was higher for the leveraged portfolio than for the long-only portfolio. A 130/30 portfolio is characterized by higher tracking errors and total risk, as measured by the portfolio's standard deviation, but the investor is compensated for bearing these risks in the higher Sharpe ratio. The execution of the McKinley Capital 130/30 portfolio has been consistent with the Guerard et al. (2009) paper in this volume.

Another risk is that the excess return can be either positive or negative. In a good year, the increased excess returns are likely. However, in a difficult year for the manager, underperformance may also be magnified. Yet another risk involves shorting stocks: demand for stock to borrow might exceed supply, or long-stock holders may require that stock be delivered out of the street name (i.e., obligate the broker or dealer to turn the certificates over to the owners). In such cases, the stock's loan cost might increase, and in extreme cases, the manager would be required to cover the stock, perhaps at an unfavorable price (a short squeeze).

There are costs associated with long–short strategies as well. First, execution costs are likely to be higher in a 130/30 format than in a long-only portfolio. Whether a manager adds more positions or takes larger positions in the existing stock, investing 160% of equity probably will cost more than investing 100% of equity, both in commissions and in market impact. Stock loan fees are another cost. While proceeds from selling short positions can be used to finance the leveraged long purchases, brokers typically charge a stock loan fee to arrange stock for short sales. The size of the fee depends on many factors including the size of the account, the number of shares to be borrowed, how and where the trade will be executed, and whether the stock is readily available to borrow or in short supply. At the low

[11] "An Assessment of a McKinley Capital Non-U.S. Developed Growth Portfolio – A Review of the Factors, Returns, and Implementation," August 2006.

end, stock borrowing fees are 25–50 basis points annualized, but at the high end these fees can reach hundreds of basis points for hard-to-borrow stocks. For legal and technical reasons, 130/30 accounts are often held in onshore or offshore limited partnerships or funds. There are audit, legal, and service fees associated with these formats. Even if the 130/30 accounts are held on a separate account basis, there are fees associated with establishing and maintaining prime brokerage accounts (as traditional custodian banks generally cannot hold short positions).

Finally, management fees are often higher for 130/30 accounts than for long-only accounts. These higher fees compensate the manager for managing more assets, both long and short, as well as for the generally greater degree of complexity of such portfolios. A prudent investor should understand and accept the higher risks and costs before undertaking a 130/30 investment.

3.6 Summary

In the 1990–2007 period, Markowitz addressed issues of using fundamental variables in stock selection, portfolio construction, and data mining tests. Extensive simulation and 130/30 research analyses were developed. Harry Markowitz continues to address problems of interest to academicians and investment professionals. The issues summarized in this chapter may been be cited in the financial literature long after 2052. If the reader is confused by that date, one might need to re-read the first three chapters.

References

Arnott, R. 1985. “The Use and Misuse of Consensus Earnings.” *Journal of Portfolio Management* 11, 18–27.

Basu, S. 1977. “Investment Performance of Common Stocks in Relations to their Price Earnings Ratios: A Test of Market Efficiency.” *Journal of Finance* 32, 663–682.

Beaton, A.E. and J.W. Tukey. 1974. “The Fitting of Power Series, Meaning Polynomials, Illustrated on Bank-Spectroscopic Data,” *Technometrics* 16, 147–185.

Black, F., M.C. Jensen, and M. Scholes. 1972. “The Capital Asset Pricing Model: Some Empirical Tests,” in *Studies in the Theory of Capital Markets*, edited by Jensen M. New York: Praeger.

Bloch, M., J.B. Guerard Jr., H.M. Markowitz, P. Todd, and G.-L. Xu. 1993. “A Comparison of Some Aspects of the U.S. and Japanese Equity Markets.” *Japan & the World Economy* 5, 3–26.

Brealey, R.A., S.C. Myers, and F. Allen. 2006. *Principles of Corporate Finance*. 8th edition. New York: McGraw-Hill/Irwin, Chapter 8.

Brown, L.D. 1999 and 2008. Annotated I/B/E/S Bibliography.

Brush, J.S. 2001. “Price Momentum: A Twenty-Year Research Effort”. Columbine Newsletter Special Issue.

Brush, J.S. 2007. “A Flexible Theory of Price Momentum”. *Journal of Investing* 16.

Brush, J.S. and K.E. Boles. 1983. “The Predictive Power in Relative Strength and CAPM”. *Journal of Portfolio Management* 9, 20–23.

Chan, L.K.C., Y. Hamao, and J. Lakonishok. 1991. “Fundamentals and Stock Returns in Japan”. *Journal of Finance* 46, 1739–1764.

Chan, L.K.C., J. Lakonishok, and Y. Hamao. 1993. “Can Fundamentals Predict Japanese Stock Returns” *Financial Analysts Journal*.

Chan, L.K.C., N. Jegadeesh, and J. Lakonishok. 1996. “Momentum Strategies.” *Journal of Finance* 51, 1681–1713.

Clarke, R.G., H. de Silva, and S. Thorley. 2002. “Portfolio Constraints and the Fundamental Law of Active Management,” *Financial Analyst Journal*, September/October, 48–66.

Clarke, R.G., H. de Silva, and S. Sapra, 2004. “Toward More Information-Efficient Portfolios,” *The Journal of Portfolio Management* 30, Fall, 54–63.

Cragg, J.G. and B.G. Malkiel. 1968. “The Consensus and Accuracy of Some Predictions in the growth of Corporate Earnings,” *Journal of Finance* 33, 67–84.

Dimson, E. 1988. *Stock Market Anomalies*. Cambridge: Cambridge University Press.

Elton, E.J., M.J. Gruber, and M. Gultekin. 1981. “Expectations and Share Prices.” *Management Science* 27, 975–987.

Fama, E.F. 1976. *Foundations of Finance*. New York: Basic Books.

Fama, E.F. and K.R. French. 1992. “Cross-Sectional Variation in Expected Stock Returns.” *Journal of Finance* 47, 427–465.

Fama, E.F. and K.R. French. 1995. “Size and the Book-to-Market Factors in Earnings and Returns. *Journal of Finance* 50, 131–155.

Fama, E.F. and K.R. French. 2008. “Dissecting Anomalies”. *Journal of Finance* 63, 1653–1678.

Fama, E.F. and J.D. MacBeth. 1973. “Risk, Return, and Equilibrium: Empirical Tests.” *Journal of Political Economy* 81, 607–636.

Farrell, J.L. Ir. 1974. “Analyzing Covariance of Returns to Determine Homogeneous Stock Groupings.” *Journal of Business* 47, 186–207.

Farrell, J.L. Ir. 1997. *Portfolio Management: Theory and Applications*. New York: McGraw-Hill/Irwin.

Fogler, R., K. John, and I. Tipton. 1981. “Three Factors, Interest Rate Differentials, and Stock Groups.” *Journal of Finance* 36, 323–335.

Givoly, D. and J. Lakonishok. 1979. “The Information Content of Financial Analysts’ Forecasts of Earnings,” *Journal of Accounting and Economics* 1, 165–185.

Graham, B., D. Dodd, and S. Cottle. 1934, 1962. *Security Analysis: Principles and Technique*. New York: McGraw-Hill Book Company. Editions one and four.

Grinhold, R. and R. Kahn. 1999. *Active Portfolio Management*. New York: McGraw-Hill/Irwin.

Guerard, J.B. Jr. 2006. “Quantitative Stock Selection in Japan and the US: Some Past and Current Issues.” *Journal of Investing*, 15.

Guerard J.B Jr and A. Mark. 2003. “The Optimization of Efficient Portfolios: The Case for a Quadratic R&D Term” *Research in Finance* 20, 213–247.

Guerard, J.B. Jr. and B.K. Stone. 1992. “Composite Forecasting of Annual Corporate Earnings.” *Research in Finance* 10. Edited by Chen A.

Guerard, J.B Jr., M. Takano, and Y. Yamane. 1993. “The Development of Efficient Portfolios in Japan with Particular Emphasis on Sales and Earnings Forecasting.” *Annals of Operations Research* 45, 91–108.

Guerard, J.B. Jr., M. Gultekin, and B.K. Stone. 1997. “The Role of Fundamental Data and Analysts’ Earnings Breadth, Forecasts, and Revisions in the Creation of Efficient Portfolios.” In *Research in Finance* 15, edited by Chen A.

Guerard, J.B., S.C. Chettiappan, and G.-L. Xu. 2009. “Stock-Selection Modeling and Data Mining Corrections: Long-Only versus 130/30 Models”. This volume.

Gunst, R.F., J.T. Webster, and R.L. Mason. 1976. “A Comparison of Least Squares and Latent Root Regression Estimators.” *Technometrics* 18, 75–83.

Haugen, R., and N. Baker. 2009. “Case Closed”. This volume.

Hawawini, G. and D.B. Keim. 1995. “On the Predictability of Common Stock Returns: World-Wide Evidence”. In Jarrow RA, Maksimovic V, and Ziemba WT, Editors, *Handbooks in Operations Research and Management Science: Finance* 9, 497–537.

Hawkins, E.H., S.C. Chamberlain, and W.E. Daniel. 1984. "Earnings Expectations and Security Prices." *Financial Analysts Journal*, 24–39.

Jacobs, B.I. and K.N. Levy. 1988. "Disentangling Equity Return and Regularities: New Insights and Investment Opportunities." *Financial Analyst Journal* May/June, 18–43.

Jacobs, B.I. and K.N. Levy. 2006. "Enhanced Active Equity Strategies," *Journal of Portfolio Management* 22, Spring, 45–55.

Jacobs, B.I., K.N. Levy, and H.M. Markowitz. 2004. "Financial Market Simulation," *Journal of Portfolio Management* 30, 142–151.

Jacobs, B.I., K.N. Levy, and H.M. Markowitz. 2005a. "Trimability and Fast Optimization of Long/Short Portfolios". *Financial Analysts Journal* 62.

Jacobs, B.I., K.N. Levy, and H.M. Markowitz, 2005b. "Portfolio Optimization with Factors, Scenarios, and Realistic Short Positions". *Operations Research* 53, 586–599.

Jacobs, B.I., K.N. Levy, and D. Starer. 1998. "On the optimality of Long-short strategies," *Financial Analysts Journal* 54, April/May. 40–51.

King, B.F. 1966. "Market and Industry Factors in Stock Price Behavior." *Journal of Business* 39, 139–191.

Korajczyk, R.A. and R. Sadka. 2004. "Are Momentum Profits Robust to Trading Costs?" *Journal of Finance* 59, 1039–1082.

Lakonishok, J., A. Shleifer, and R.W. Vishny. 1994. "Contrarian Investment, Extrapolation and Risk." *Journal of Finance* 49, 1541–1578.

Lesmond, D.A., M.J. Schill, and C. Zhou. 2004. "The Illusory Nature of Trading Profits". *Journal of Financial Economics* 71, 349–380.

Levy, H. 1999. Introduction to Investments. Cincinnati: South-Western College Publishing. Second Edition.

Lintner, J. 1965a. "Security Prices, Risk, and the Maximum Gain from Diversification." *Journal of Finance* 30.

Lintner, J. 1965b. "The Valuation of Risk Assets on the Selection of Risky Investments in Stock Portfolios and Capital Investments." *The Review of Economics and Statistics* 57, 13–37.

Malkiel, B. 1996. *A Random Walk Down Wall Street*, New York: W.W. Norton & Company. Sixth Edition.

Markowitz, H.M. 1952. "Portfolio Selection." *Journal of Finance* 7, 77–91.

Markowitz, H.M. 1959. *Portfolio Selection: Efficient Diversification of Investment*. Cowles Foundation Monograph No.16. New York, Wiley & Sons.

Markowitz, H.M. 1976. "Investment in the Long Run: New Evidence for an Old Rule," *Journal of Finance* 31, 1273–1286.

Markowitz, H.M. and E.L. Xu. 1994. "Data Mining Corrections." *Journal of Portfolio Management* 19, Fall. 60–69.

Markowitz, H.M. 2000. *Mean-Variance Analysis in Portfolio Choice and Capital Markets*. New Hope, PA: Frank J. Fabozzi Associates.

Markowitz, H.M. 2002. Efficient portfolios, sparse matrices, and entities: A retrospective. *Operations Research* 50(1), 154–160.

Markowitz, H.M., B. Hausner, and H.W. Karr. 1963. *SIMSCRIPT: A Simulation Programming Language*. Englewood Cliffs, NJ: Prentice-Hall Book Company.

Middleton, P. 2007. "130/30 Portfolios – Here Comes the Wave," *Merrill Lynch, Global Asset Managers*, Industry Overview, March 6.

Mossin, J. 1966. "Equilibrium in a Capital Asset Market." *Econometrica* 34, 768–783.

Mossin, J. 1973. *Theory of Financial Markets*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Nerlove, M. 1968. "Factors Affecting Differences Among Rates of Return on Investments in Individual Common Stocks". *Review of Economics and Statistics* 50, 312–331.

Park, S. and S. Bucknor. 2006. "120/20 Portfolios: Long-Only Managers' Way of Reaching for Higher Alpha," CRA RogersCasey, *U.S. Equity Research*, November.

Ramnath, S., S. Rock, and P. Shane. 2008. "The Financial Analyst Forecasting Literature: A Taxonomy with Suggestions for Further Research. *International Journal of Forecasting* 24, 34–75.

Reinganum, M. 1981. “The Arbitrage Pricing Theory: Some Empirical Tests.” *Journal of Finance* 36, 313–321.

Rosenberg, B. 1974. “Extra-Market Components of Covariance in Security Returns.” *Journal of Financial and Quantitative Analysis* 9, 263–274.

Rosenberg, B. and V. Marathe, 1979. “Tests of Capital Asset Pricing Hypotheses.” In *Research in Finance* 1,1979, edited by Levy H.

Rudd, A., and B. Rosenberg. 1980. “The ‘Market Model’ in Investment Management.” *Journal of Finance* 35, 597–607.

Rudd, A, and B. Rosenberg. 1979. “Realistic Portfolio Optimization.” In *Portfolio Theory, 25 Years After*. Edited by Elton E, and Gruber M Amsterdam: North-Holland.

Sharpe, W.F. 1963. “A Simplified Model for Portfolio Analysis.” *Management Science* 9, 277–293.

Sharpe, W.F. 1964. “Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk.” *Journal of Finance* 19, 425–442.

Sharpe, W.F. 1971. “A Linear Programming Approximation for the General Portfolio Analysis Problem.” *Journal of Financial and Quantitative Analysis 6*.

Stone, B.K. 1970. *Risk, Return, and Equilibrium: A General Single-Period Theory of Asset Selection and Capital Market Equilibrium*. Cambridge, MA: MIT Press.

Stone, B.K. 1973. “A Linear Programming Formulation of the General Portfolio Selection Problem.” *Journal of Financial and Quantitative Analysis* 8, 621–636.

Wheeler, L.B. 1994. “Changes in Consensus Earnings Estimates and Their Impact on Stock Returns.” In *The Handbook of Corporate Earnings Analysis*. Edited by Bruce B, and Epstein CB.

Williams, J.B. 1938. *The Theory of Investment Value*. Cambridge: Harvard University Press.

Von Hohenbalken, B. 1975. “A Finite Algorithm to Maximize Certain Pseudoconcave Functions on Polytopes.” *Mathematical Programming*.

Ziemba, W.T. 1990. “Fundamental Factors in US and Japanese Stock Returns.” Berkeley Program in Finance.

Ziemba, W.T. 1992. *Invest Japan*. Chicago: Probus Publishing Co.

Chapter 4
Markowitz's Mean–Variance Rule and the Talmudic Diversification Recommendation

Haim Levy and Ran Duchin

> *"A man should always place his money, one-third into land, a third into merchandise and keep a third in hand" (Babylonian Talmud)*

This 1500-year-old investment advice recommends a naïve diversification strategy, or a "1/3 rule," which generalizes to the "$1/N$ rule" with N assets [see Levy and Sarnat (1972)]. This diversification strategy contradicts the diversification policy implied by von-Neumann and Morgenstrern (1953) expected utility theory (EU) in general, and Markowitz's (1952a, b) Mean–Variance (M–V) implied diversification in particular. Obviously, it also contradicts Sharpe and Lintner's CAPM [see Sharpe (1964) and Lintner (1965)]. The naïve diversification policy called hereafter the "$1/N$ rule" ignores the distribution of rates of return, and in particular means, variances and correlations. Surprisingly, experimental studies reveal that individual investors behave according to the "$1/N$ rule" much more than according to expected utility theory. For example, Kroll, Levy, and Rapoport (1988) show that when the subjects face three assets A, B, and C, changing the correlation between B and C from $\delta_{\mathrm{BC}} = 0$ to $\delta_{\mathrm{BC}} = 0.8$ or even $\delta_{\mathrm{BC}} = -0.8$ does not change much the diversification policy chosen by the subjects, which contradicts EU as well as M–V theories. In a more recent study, Benartzi and Thaler (B&T) (2001) find that subjects tend to follow the "$1/N$ rule" regardless of the characteristics of the N available funds.[1]

Does ignoring asset parameters, and in particular their correlations, fit into the category of psychological errors – a subject of increasing popularity in behavioral finance and behavioral economics studies? Namely, did the investors and subjects who participated in the KL&R and B&T experiments simply adopt a wrong invest-

H. Levy (✉)
Jerusalem School of Business, The Hebrew University, Jerusalem 91905, Israel
e-mail: mshlevy@mscc.huji.ac.il

R. Duchin
University of Michigan Ross School of Business, 701 Tappan St., Ann Arbor, MI 48109, USA
e-mail: duchin@bus.umich.edu

[1] The above experiment is conducted with a relatively small N, hence, its results only apply to relatively small values of N. We believe no one would agree that investors who face say 5,000 assets would invest 1/5,000 in each asset. However, in say, 4 out of the 5,000 assets which one selects, the investment would be $\frac{1}{4}$ in each asset.

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_4,
© Springer Science+Business Media, LLC 2010

ment strategy and hence suffered a substantial loss in terms of expected utility? Were they behaving irrationally? In this study, we address these questions taking into account the different sets of preferences and available assets faced by different types of investors. As we shall see, the number of assets in the portfolio and the risk preferences play a key role in measuring the loss involved with the "$1/N$ rule."

While at a first glance the "$1/N$ rule" seems to always be inferior to the classical EU and M–V diversification strategies, we will show that this is not always true. This is because the "$1/N$ rule" has one important advantage over classical diversification methods: it is not exposed to estimation errors that cause investors who follow the EU or M–V rules to either over-invest or under-invest in a given security. Thus, the "$1/N$ rule" may actually outperform the classical diversification methods in the *out-of-sample* framework.

To measure the loss due to the employment of the "$1/N$ rule," we consider three scenarios:

(a) *Full information*. Investors have *precise knowledge* of the future joint distribution of returns; hence there are no estimation errors and employing the "$1/N$ rule" is irrational and incurs an expected utility loss by definition. As investors observe the *ex ante* distribution, this extreme case assumes that the ex post distribution is *identical* to the *ex ante* distribution. In this case, the appropriate way to measure the loss is an *in-sample* analysis, which assumes a *stable and fully known* distribution of returns.
(b) *Zero information*. In the other extreme case, investors have *zero knowledge* about the future distribution. Here, investing $1/N$ in each available asset is optimal.
(c) *Partial information*. The most realistic case lies between the above two extreme scenarios. Investors observe the ex post joint distribution of returns, but are aware of the fact that this distribution is not necessarily stable over time and may contain outliers. However, they are unable to predict changes in the distribution over time, hence must rely solely on the ex post distribution. In this case, the performance of the "$1/N$ rule" relative to the EU or M–V strategies depends on the degree of stability of the distribution over time. If there is a substantial and unpredictable instability, the "$1/N$ rule" may induce a very small loss or even a gain relative to the EU or M–V strategies, because the latter rely on the assumption that the joint distribution is stable over time.

To analyze the "$1/N$ rule" investment strategy vs. the EU and M–V strategies, we conduct an *out-of-sample* analysis. If the instability of the joint distribution is extreme (and unpredictable) such that the ex post distribution is of no value at all, we expect the "$1/N$ rule" to be optimal. If the distribution is completely stable then the ex post parameters can be usefully employed, and we expect the EU or M–V strategies to be optimal. In practice, distributions lie between the two extremes and measuring the loss/gain induced by the "$1/N$ rule" becomes an empirical task. Moreover, as some assets are more stable than others, the "$1/N$ rule" might be superior for some assets, while the EU or M–V rules might be superior for others. Yet, we should emphasize that the M–V optimal portfolio is very sensitive to changes in the parameters, and this gives an advantage to the "$1/N$ rule" in the out-of-sample analysis [for studies of the sensitivity of the M–V portfolio mix to changes in the parameters, [see Best and Grauer (1991), Green and Hollifield (1992), and Jagannathan and Ma (2003)].

Furthermore, we structure our empirical investigation to distinguish between the performance of the "$1/N$ rule" associated with small and large portfolios. The reason is that there is a sharp difference between the number of assets held by institutions and individual investors, a factor which is relevant, as we shall see below, to the loss induced by the employment of the "$1/N$ rule."[2] Empirical findings reveal that household investors tend to hold a relatively small number of assets in their portfolio [see Blume et al. (1974), Blume and Friend (1975), Levy (2006), and Barber and Odean (1999)]. This policy may be due to fixed transaction costs, asymmetry of information, and many other reasons [see Levy (1978) and Merton (1987)]. We show that the Talmudic "$1/N$ rule" incurs in some (albeit not all) cases a loss to both institutional investors and households. However, the loss is much smaller for households, since their portfolios contain a relatively small number of assets. Moreover, the "$1/N$ rule" may even induce a gain over the M–V portfolio in practice (out-of-sample). As mentioned above, the explanation for this result is that the "$1/N$ rule" contains a hedge against estimation errors induced by the instability of the the empirical distribution of returns. Thus, employing the "$1/N$ rule" may be justified in practice.[3]

Is there a way to make use of the information content of the ex post distribution of returns and at the same time limit the exposure to estimation risk? We follow Frost and Savarino (1988), who suggest that an investment strategy that relies on constrained ex post optimization might be optimal. In particular, we implement ex post EU and M–V optimizations with constraints on the proportion invested in each asset, and use the resulting portfolio allocation out-of-sample. This policy mitigates possible extreme differences between the true (*ex ante*) distribution of returns and the observed (ex post) distribution of returns. Hence, it has some degree of the hedge against estimation errors, which reduces the hedging advantage of the "$1/N$ rule" relative to the EU or M–V strategies.

The structure of the paper is as follows: In Sect. 4.1, we present the *in-sample* and *out-of-sample* methodologies to measure the loss of investors who employ the "$1/N$ rule" rather than the optimal EU or M–V investment strategies. We compare the "$1/N$ rule" first to Markowitz's M–V rule and then to expected utility maximiza-

---

[2] When short sales are not allowed, the EU and M–V strategies are very different from the "$1/N$ rule" and the difference increases with the number of available assets. The reason is that for large values of N, the EU and M–V strategies generally advise investors to hold zero proportions in most assets [see Levy (1983)], while the "$1/N$ rule" advises positive proportions in all assets by definition.

[3] Markowitz, the "father" of portfolio selection himself, reports that he uses the same rule and justifies it by minimizing the future regret. For further reference, see Zweig (1998) and Benartzi and Thaler (2001).

tion with various relative risk-aversion coefficients. Section 4.2 provides the results. Concluding remarks are given in Sect. 4.3.

4.1 The Methodology

4.1.1 Mean–Variance Analysis

We start our empirical investigation by employing the well-known M–V approximation to any concave preference [see Levy and Markowitz (1979) and Markowitz (1991)], and measure the loss induced by employing the "$1/N$ rule" in the M–V framework. To be more specific, Levy and Markowitz show that if returns fall in the range [−30%, +60%], the M–V rule (or quadratic preferences) provides an excellent approximation to the most commonly employed risk-aversion preferences [see Levy and Markowitz (1979), p. 311].[4] As in our study most rates of *return fall* in this range, we can safely use the M–V rule. Thus, we conduct *in-sample* and *out-of-sample* M–V efficiency analysis to find the optimal M–V portfolio of risky assets, as follows:

$$\begin{aligned} &\text{Min}\ \underline{x}^{\mathrm{T}} V \underline{x} \\ &\text{s.t.}\ \underline{x} \geq 0 \\ &\underline{x}^{\mathrm{T}} \underline{\overline{R}} + (1 - \underline{x}^{\mathrm{T}} \underline{1}) r = \overline{R}_{\mathrm{p}}, \end{aligned} \tag{4.1}$$

where V is the variance–covariance matrix, $\bar{R}_i$ is the sample mean rate of return on the ith asset, $\bar{R}_{\mathrm{p}}$ is the portfolio mean rate of return, and r is the riskless interest rate. For the "$1/N$" portfolio, we have

$$\begin{aligned} &\text{Portfolio Variance} = \underline{P}^{\mathrm{T}} V \underline{P}, \\ &\text{Portfolio Mean} = \underline{P}^{\mathrm{T}} \underline{\overline{R}} + (1 - \underline{P}^{\mathrm{T}} \underline{1}) r = \overline{R}_{\mathrm{p}}, \end{aligned} \tag{4.2}$$

where $P^{\mathrm{T}} = (1/N,\ 1/N, \ldots, 1/N)$, and, of course, no optimization is involved in this case. As we shall demonstrate with the data, the loss induced by the "$1/N$ rule" is measured with *in-and-out-of-sample* data by the *vertical* distance between the optimal M–V portfolio and the "$1/N$ rule" portfolio. Thus, for a given risk (i.e., standard deviation), the loss is measured in terms of expected rate of return lost by the employment of the "$1/N$ rule." For simplicity, we demonstrate the loss

[4] Note that the M–V criterion assumes either quadratic preferences or concave preferences with an Elliptic distribution (e.g., a Normal distribution). Thus, preferences need to be concave. It is interesting to note that in the same year that Markowitz published his seminal M–V paper, 1952, he also published another paper which is less famous but not less important, which assumes a non-concave utility function [see Markowitz (1952b)]. In that paper, there are a few elements of Prospect Theory, which was only published in 1979!

associated with the "$1/N$ rule" portfolio with zero investment in the riskless asset. Thus, the loss is measured by the vertical line arising from the "$1/N$ portfolio" to the efficient frontier, on which the optimal portfolios lie.

4.1.2 Expected Utility Analysis

As the M–V rule is not always applicable (e.g., when distributions are highly skewed), we also analyze the efficiency of the "$1/N$ rule" in expected utility framework. Suppose that $\underline{R}$ is a vector of returns whose joint density distribution is $f(\underline{R})$ corresponding to N assets. A portfolio is defined by $\underline{x}$, where $\underline{x}$ is a vector of investment proportions. We obtain the optimal investment proportions, which maximize the expected utility, by solving the following problem:

$$\begin{aligned} &\underset{\underline{x}}{\mathrm{Max}} \int_{\underline{R}} U(\underline{x}'\underline{R}) f(\underline{R}) \mathrm{d}\underline{R} \\ &\text{s.t.} \\ &\underline{x} \geq 0 \\ &\underline{x}'\underline{1} = 1, \end{aligned} \tag{4.3}$$

where $U(\cdot)$ is a utility function. Solving the optimization problem given above for a vector $\underline{x}^*$, the expected utility due to the optimal investment policy can be calculated. We denote the maximal expected utility given in (4.3) by $\mathrm{EU}_{\max}(\cdot)$. We measure the loss due to the employment of the "$1/N$ rule" in dollars and not in utility terms (which is affected by a linear transformation). Therefore, we calculate the Certainty Equivalent (CE) induced by each investment strategy, given by:

$$\mathrm{EU}_{\max}(\cdot) = U(\mathrm{CE}_{\max}). \tag{4.4}$$

In other words, we find the certainty equivalent corresponding to the investment policy which maximized expected utility. Denoting the number of available assets by N, one can calculate the expected utility induced by the "$1/N$ rule" for the same distributions of historical returns, as follows:

$$\mathrm{EU}_{1/N}(\cdot) = \int_{\underline{R}} U\left(\frac{\underline{1}'\underline{R}}{N}\right) f(\underline{R}) \mathrm{d}\underline{R}. \tag{4.5}$$

Once again, we find by the CE corresponding to this policy, denoted by $\mathrm{CE}_{1/N}$, as follows

$$\mathrm{EU}_{1/N}(\cdot) = U(\mathrm{CE}_{1/N}). \tag{4.6}$$

We use the difference between CE_{max} and $CE_{1/N}$ to measure the induced loss/gain due to the employment of the "$1/N$ rule" in dollar terms.[5]

4.1.2.1 In-Sample Analysis: Stable Distributions Over Time

The *in-sample* loss in the expected utility framework is measured as follows: Suppose that one faces N assets with T rates of return corresponding to each asset. This is the multivariate empirical distribution. As we have no knowledge regarding the future multivariate distribution, we substitute in (4.3) and (4.5) the empirical $f(\underline{R})$ and solve it for the expected utility as given by these two equations. Then we calculate the CE loss [see (4.4) and (4.6)] for a given utility function. Note that in *in-sample*, we must have: $CE_{1/N} < CE_{max}$.

4.1.2.2 Out-of-Sample Analysis: Unstable Distributions Over Time

While the "1/*N rule*" induces an *in-sample* loss by definition, this is not necessarily so with the *out-of-sample* analysis (which is the relevant one for investors), because distributions are generally not stable. Let us elaborate on the *out-of-sample* technique employed in this study. We divide the T periods into two subperiods: period $1, 2, \ldots, T_1$ (the ex post period) and period $T_1 + 1, T_1 + 2, \ldots, T_1$, the *ex ante* period, which is used to obtain the *out-of-sample* returns. We record the two returns $R_{T_1+1}(1/n) \equiv R_{T_1+1}$ and $R_{T_1+1}(\mathrm{EU}_{max}(\cdot)) \equiv R^*_{T_1+1}$, where R_{T_1+1} is the return in period T_1+1 on the "$1/N$ rule" portfolio and $R^*_{T_1+1}$ is the return on the $\mathrm{EU}_{max}(\cdot)$ portfolio whose weights were calculated based on the period $1, 2, \ldots, T_1$. Namely, we have $R_{T_1+1} = \sum_{i=1}^{n} \frac{1}{n} R_{i,T+1}$ and $R^*_{T_1+1} = \sum_{i=1}^{n} x_i^* R_{i,T+1}$, where x_i^* is the vector of investment proportions obtained by solving (4.3), based on $f(\underline{R})$ as estimated according to the first T_1 periods. We repeat the procedure by deleting the first observations from T_1 (from all available assets) and adding the observations T_1+1 to the sample. Using the historical returns, from period 2 to period $T_1 + 1$, we repeat the same procedure discussed above to obtain R_{T_1+2} and $R^*_{T_1+2}$. Continuing the procedure, we obtain two *out-of-sample* vectors of returns $\underline{R} = R_{T_1+1}, R_{T_1+2}, \ldots, R_T$ and $\underline{R}^* = R^*_{T_1+1}, R^*_{T_1+2}, \ldots, R^*_T$ corresponding to the two investment strategies under consideration.

Now, suppose our investor faces these two out-of-sample vectors of returns $\underline{R}$ and $\underline{R}^*$ and wishes to invest for one period. Which vector is preferred? The question

[5] Kroll et al. (1984) measure the expected utility loss using the index I, which is given by:

$$0 \le 1 = \frac{\mathrm{EU}_{1/N}(\cdot) - \mathrm{EU}^*(\cdot)}{\mathrm{EU}_{max}(\cdot) - \mathrm{EU}^*(\cdot)} \le 1,$$

where $\mathrm{EU}^*()$ is the expected utility of an arbitrary portfolio. We believe the CE is a better measure because it tells us the loss in dollar amounts, while an index of say $I = 0.9$ has a less intuitive interpretation.

boils down to the following question: Suppose that an investor is given the choice to draw one observation at random from either $\underline{R}$ or $\underline{R}^*$. Which is a better choice? Of course, with stable distribution and with a large number of periods, $\underline{R}^*$ must yield a higher expected utility than $\underline{R}$. However, this is not the case with *unstable* distributions over time. To answer this question, we calculate the CE [see (4.4) and (4.6)] for these two vectors. If $CE_{1/N} > CE_{max}$, the "$1/N$ rule" outperforms the expected utility maximization strategy. The opposite conclusion holds if $CE_{1/N} < CE_{max}$.

Relying on ex post data implies an overinvestment in securities with a favorable estimation error and underinvestment in securities with an unfavorable error. In particular, the optimal investment error is sensitive to the difference between the true mean μ_i, and the sample estimation of μ_i given by $\bar{R}_i$ [see Best and Grauer (1991)].[6] This gives an edge to the "$1/N$ rule" since it is not exposed to these errors. In other words, the sample distribution contains errors that may affect the optimal investment proportions. There are a few methods for minimizing these possible errors. Frost and Savarino (1986) suggest a Bayesian framework in portfolio selection. In a follow-up paper (1988), they suggest to impose a constraint on the maximum investment proportion in each asset. Thus, if an error exists due to the difference between $\bar{R}_i$ and μ_i, e.g., $\bar{R}_i > \mu_i$, the constraint on the investment proportions mitigates this error as it prevents investing a very large proportion in the ith asset. As mentioned above, the "$1/N$ rule" strategy is not affected by the above-mentioned sample errors. Therefore, apart from comparing the max EU($\cdot$) strategy with the "$1/N$ rule" strategy, we also compare the "$1/N$ rule" with the max EU($\cdot$) strategy imposing constraints of the form $x_i \leq p$ on the maximal investment proportions allowed in any particular asset i, where p is some given proportion. Using ex post data, a maximization of EU($\cdot$) with constraints may prove to be better than the maximization of EU($\cdot$) without constraints, hence, may be the optimal investment policy in the *out-of-sample* analysis. Note, however, that imposing strict constraints simply coincides with the "$1/N$ rule." For example, with $N=20$, if $x_i \leq 5\%$ we obtain the "$1/N$ rule." Thus, the expected utility maximization with constraints is a compromise between the pure expected utility maximization strategy (with ineffective constraints) and the "$1/N$ rule" strategy (with very strict constraints).

4.2 Data and Results

4.2.1 The Data

Our data comprise of monthly value-weighted returns on the 30 Fama-French industry portfolios (for details, see: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html). We focus on the period between January 1996 and May

[6] On the same paradoxical results of the M–V portfolios which are based on ex post parameters, see Levy (1983), Green and Hollifield (1992), and Jagannathan and Ma (2003).

2007, with a total of 137 monthly observations for each portfolio.[7] Table 4.1 reports summary statistics for the monthly returns on the 30 industry portfolios, and shows that the average return was approximately 1%, with a standard deviation of about 4–10%. The table indicates that there is a significant amount of cross-sectional variation between the portfolios in terms of average returns and standard deviations, which implies that they are not perfect substitutes. The last two columns report the

Table 4.1 Summary statistics: monthly returns (%) on the 30 Fama-French industry portfolios

Industry portfolio	Expected return	Standard deviation	Min return	Max return
Food products	1.00	4.86	−28.26	33.48
Beer & liquor	1.25	7.52	−29.19	89.18
Tobacco products	1.17	5.89	−24.93	33.30
Recreation	1.15	8.93	−44.45	66.81
Printing and publishing	1.03	6.97	−30.45	53.43
Consumer goods	0.92	5.60	−33.65	52.20
Apparel	1.10	6.97	−31.42	54.18
Healthcare, medical equipment, pharm	1.10	5.78	-34.74	38.66
Chemicals	1.07	6.26	−33.31	46.99
Textiles	0.98	7.63	−32.40	57.05
Construction and construction materials	0.98	6.79	−31.81	43.28
Steel works, etc.	1.02	8.35	−31.07	80.72
Fabricated products and machinery	1.09	7.19	−33.38	51.94
Electrical equipment	1.23	7.69	−34.50	59.62
Automobiles and trucks	1.14	7.84	−34.90	81.93
Aircraft, ships, and railroad equipment	1.14	7.77	−30.86	49.60
Precious metals, metal mining	0.98	7.00	−33.16	45.57
Coal	1.25	8.81	−30.11	77.54
Petroleum and natural gas	1.11	6.08	−29.72	39.16
Utilities	0.93	5.67	−29.84	43.16
Communication	0.88	4.60	−21.56	28.16
Personal and business services	1.23	8.72	−50.85	73.32
Business equipment	1.14	6.84	−31.65	38.28
Business supplies and shipping containers	1.09	5.91	−29.30	43.38
Transportation	0.94	7.23	−34.52	65.35
Wholesale	0.83	7.56	−44.47	59.21
Retail	1.03	6.04	−30.25	37.75
Restaurants, hotels, motels	1.12	6.96	−31.35	31.49
Banking, insurance, real estate, trading	1.10	6.80	−39.47	59.85
Everything else	0.80	6.81	−31.80	45.99

This table presents average monthly returns, monthly return volatilities, and minimum and maximum monthly returns for the 30 Fama-French value-weighted industry portfolios, for the period January 1996 to May 2007

[7] We repeated our analysis for other periods and investment horizons and obtained qualitatively similar results.

minimum and maximum monthly returns from January 1996 to May 2007, and show that the returns on most portfolios were confined between −30% and +60%, which validates the M–V approximation, as noted above.

Table 4.2 reports the correlations between the monthly returns on the 30 portfolios for January 1996 to May 2007. As the table shows, all portfolios are positively correlated (probably due to their mutual co-movement with the market portfolio), with an average correlation of 0.65. Thus, we conclude that our assets are far from being perfectly correlated, and therefore there is room for gain from diversification.

4.2.2 Mean–Variance Results

Our first task is to estimate the loss in terms of expected return induced by following the "$1/N$" rule instead of the M–V strategy, abstracting away from estimation error effects. Thus, we assume that the empirical ex post distribution of returns is the true *ex ante* distribution of returns, and estimate the *in-sample* "added" return earned by the M–V efficient portfolio that has the same standard deviation as the "$1/N$" portfolio.

Table 4.3 reports the results of the *in-sample* analysis for selected values of N ranging from $N=5$ to $N=30$. Note that all numbers in this and subsequent tables are given in percents. Thus, following the "$1/N$ rule" yields a loss of 0.062% per month when $N=5$, 0.13% per month when $N=10$, and so on until we reach $N=30$, in which case the loss associated with the "$1/N$ rule" is 0.444% per month. These results bring about two important conclusions. (1) In *in-sample*, the "$1/N$ rule" induces significant losses compared to the M–V strategy (Recall that the average monthly return is approximately 1%. Thus, a loss of 0.318% for $N=20$ comprises almost one-third of the average monthly return.) and (2) these losses grow significantly when the number of available assets (N) grows. As noted above, the relative performance of the "$1/N$ rule" worsens as N increases because the optimal M–V strategy tends to invest in a small number of assets.

Figures 4.1 and 4.2 visually illustrate the in-sample loss associated with the "$1/N$ rule" for $N = 10$ and $N = 30$. In both figures, the X symbol under the M–V efficient frontier denotes the combination of the average return and the standard deviation of the "$1/N$ rule" portfolio (the naïve portfolio). To estimate the loss incurred by the "$1/N$ rule," one has to measure the vertical distance between the X and the portfolio on the efficient frontier that has the same standard deviation as the "$1/N$ rule" portfolio. The figures clearly illustrate that the losses are substantial and increase as the number of assets increases.

Having found evidence consistent with substantial *in-sample* losses induced by the "$1/N$" rule, we turn to the *out-of-sample* analysis, which is the most relevant from a practical point of view. We divide the sample period in the middle, with the first half deemed as the *in-sample* period used to obtain the (ex post) M–V-efficient portfolio allocations, and the second half deemed as the *out-of-sample* (*ex ante*) period used to evaluate the performance of the investment diversification strategy

Table 4.2 Correlations between monthly returns on the 30 Fama-French industry portfolios: January 1996 to May 2007

1.00																													
0.70	1.00																												
0.67	0.45	1.00																											
0.72	0.67	0.50	1.00																										
0.75	0.62	0.50	0.73	1.00																									
0.78	0.65	0.54	0.69	0.72	1.00																								
0.71	0.60	0.49	0.72	0.73	0.67	1.00																							
0.78	0.64	0.57	0.69	0.67	0.74	0.63	1.00																						
0.75	0.63	0.51	0.72	0.71	0.75	0.70	0.71	1.00																					
0.72	0.68	0.47	0.76	0.77	0.70	0.78	0.63	0.75	1.00																				
0.79	0.71	0.56	0.80	0.80	0.77	0.79	0.71	0.84	0.83	1.00																			
0.65	0.56	0.47	0.72	0.71	0.65	0.68	0.61	0.80	0.75	0.81	1.00																		
0.73	0.65	0.50	0.78	0.79	0.74	0.75	0.69	0.85	0.81	0.88	0.88	1.00																	
0.76	0.63	0.52	0.77	0.75	0.75	0.69	0.73	0.83	0.74	0.84	0.80	0.86	1.00																
0.68	0.58	0.48	0.72	0.74	0.71	0.70	0.62	0.80	0.78	0.80	0.78	0.81	0.78	1.00															
0.71	0.63	0.49	0.72	0.73	0.69	0.72	0.65	0.75	0.75	0.81	0.78	0.83	0.79	0.73	1.00														
0.55	0.49	0.40	0.60	0.57	0.53	0.55	0.51	0.66	0.61	0.68	0.73	0.71	0.63	0.59	0.61	1.00													
0.37	0.35	0.27	0.39	0.39	0.32	0.41	0.35	0.43	0.42	0.48	0.47	0.51	0.41	0.34	0.47	0.44	1.00												
0.61	0.51	0.44	0.58	0.59	0.56	0.56	0.57	0.69	0.59	0.69	0.68	0.71	0.64	0.60	0.65	0.64	0.48	1.00											
0.71	0.57	0.51	0.63	0.65	0.62	0.57	0.63	0.66	0.64	0.70	0.59	0.66	0.70	0.62	0.63	0.52	0.36	0.62	1.00										
0.65	0.51	0.45	0.62	0.61	0.59	0.55	0.60	0.65	0.57	0.64	0.59	0.64	0.65	0.60	0.56	0.46	0.30	0.49	0.63	1.00									
0.40	0.34	0.29	0.41	0.48	0.40	0.41	0.40	0.38	0.37	0.46	0.38	0.45	0.41	0.36	0.39	0.32	0.29	0.35	0.33	0.43	1.00								
0.63	0.54	0.44	0.72	0.65	0.68	0.63	0.68	0.72	0.64	0.75	0.73	0.79	0.77	0.68	0.66	0.55	0.38	0.55	0.52	0.64	0.52	1.00							
0.75	0.62	0.54	0.70	0.74	0.77	0.72	0.72	0.86	0.74	0.82	0.78	0.83	0.80	0.77	0.74	0.63	0.44	0.65	0.64	0.61	0.43	0.72	1.00						
0.72	0.61	0.49	0.75	0.77	0.67	0.74	0.65	0.77	0.79	0.82	0.80	0.84	0.79	0.74	0.82	0.64	0.50	0.63	0.65	0.62	0.40	0.66	0.76	1.00					
0.70	0.68	0.50	0.75	0.72	0.69	0.70	0.66	0.69	0.75	0.80	0.68	0.74	0.70	0.67	0.70	0.58	0.40	0.57	0.62	0.57	0.41	0.64	0.69	0.72	1.00				
0.81	0.65	0.53	0.76	0.77	0.76	0.79	0.72	0.76	0.77	0.81	0.69	0.77	0.78	0.76	0.70	0.55	0.34	0.56	0.63	0.66	0.45	0.72	0.76	0.73	0.74	1.00			
0.71	0.62	0.48	0.73	0.70	0.67	0.70	0.64	0.66	0.69	0.75	0.60	0.70	0.69	0.63	0.68	0.53	0.38	0.52	0.58	0.56	0.45	0.62	0.67	0.70	0.68	0.75	1.00		
0.82	0.69	0.56	0.80	0.78	0.75	0.75	0.76	0.80	0.78	0.86	0.76	0.83	0.84	0.77	0.78	0.63	0.42	0.68	0.80	0.71	0.42	0.71	0.80	0.82	0.75	0.81	0.72	1.00	
0.73	0.63	0.54	0.77	0.74	0.73	0.73	0.69	0.76	0.76	0.82	0.73	0.79	0.76	0.73	0.74	0.62	0.39	0.62	0.63	0.61	0.43	0.70	0.76	0.74	0.73	0.75	0.68	0.80	1.00

Table 4.3 In-sample analysis of the mean–variance approach

Number of assets (N)	Standard deviation (1)	Naïve average return (2)	M–V-efficient average return (3)	Difference = (M–V) − naïve [(4) = (3) − (2)]
5	3.830	1.020	1.082	0.062
10	3.596	0.954	1.085	0.131
15	4.056	1.017	1.244	0.227
17	4.100	1.051	1.302	0.251
20	4.048	1.161	1.479	0.318
22	4.058	1.127	1.482	0.355
25	4.119	1.109	1.503	0.394
27	4.080	1.098	1.527	0.429
30	4.040	1.071	1.515	0.444

This table reports the in-sample loss induced by following the naïve ($1/N$) investment rule rather than the optimal mean–variance investment rule. For each number of available assets, the table reports the standard deviation and the average return on the naïve portfolio, as well as the average return of the mean–variance efficient portfolio that has the same standard deviation. All numbers are reported as percents. The data consist of monthly returns on the 30 Fama-French value-weighted industry portfolios from January 1996 to May 2007

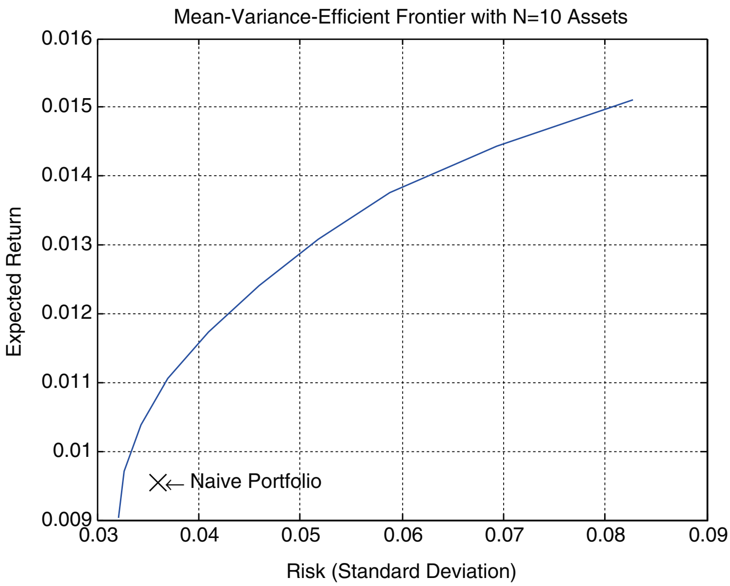


Fig. 4.1 *In-sample* mean–variance analysis of the "$1/N$ rule" when $N = 10$. This figure shows the *in-sample* performance of the "$1/N$ rule" (naïve portfolio) performance relative to the mean–variance efficient frontier when the number of available assets is 10 (the first ten assets appearing in Table 4.1). Data consist of monthly returns on the 30 Fama-French industry performance from January 1996 to May 2007

which is based on the first period data. We compare the second period M–V efficient portfolio with the performance of the "$1/N$ rule" investment strategy and with the optimal ex post M–V efficient investment strategy. However, recall that the second

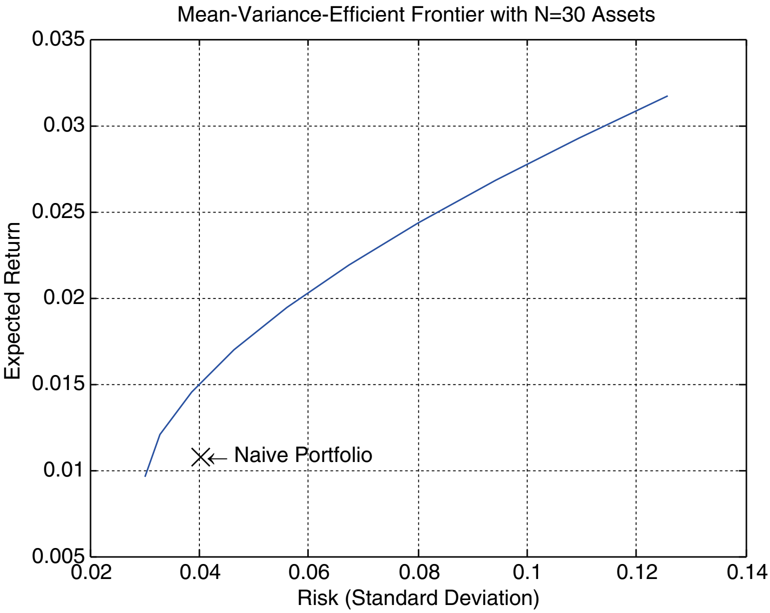


Fig. 4.2 *In-sample* mean–variance analysis of the "$1/N$ rule" when $N = 30$. This figure shows the *in-sample* performance of the "$1/N$ rule" (naïve portfolio) performance relative to the mean–variance efficient frontier when the number of available assets is 30. Data consist of monthly returns on the 30 Fama-French industry performance from January 1996 to May 2007

period M–V efficient portfolio is not available and is given just to see how the "$1/N$ rule" portfolio and the ex post M–V portfolio perform relative to this optimal *ex ante* portfolio. Table 4.4 reports the results of the *out-of-sample* analysis without imposing constraints on the proportion allocated to each asset, while Table 4.5 reports the results when the investment proportions are constrained. Both tables report the average returns corresponding to the second period: the *ex ante* M–V efficient return (i.e., the optimal investment's return based on the second half of the sample) which is unavailable and given just to measure the performance of the other two strategies, the *ex ante* M–V return (i.e., the return in the second period associated with the M–V efficient investment proportions derived from the first period), and the "$1/N$ rule" return.

Starting with Table 4.4, our results indicate that while both the *ex ante* M–V strategy and "$1/N$ rule" strategy induce significant losses relative to the optimal *ex ante* M–V strategy (which is not available as the *ex ante* parameters are unknown), the "$1/N$ rule" is superior to the M–V strategy in most cases. Again, its superiority stems from its built-in hedge against estimation errors, which are particularly pronounced when we use first half of the sample to predict parameters of the second half of the sample. However, the relative superiority of the "$1/N$ rule" gradually loses its power as the number of available assets increases. In fact, when $N = 30$, the M–V strategy actually outperforms the "$1/N$ rule." Figures 4.3 and 4.4 illustrate

Table 4.4 Out-of-sample analysis of the mean–variance approach

Number of assets (N)	Standard deviation (1)	M–V *ex ante*[a] efficient average return (2)	Ex post M–V average return (3)	Naïve average return (4)	Difference = ex post − naïve [(5) = (3) − (4)]
15	3.912	1.442	0.599	1.071	−0.472
17	3.947	1.606	0.609	1.148	−0.539
20	3.998	1.679	0.999	1.255	−0.256
22	3.998	1.679	0.999	1.185	−0.186
25	4.060	1.700	1.010	1.146	−0.136
27	3.997	1.679	1.109	1.111	−0.002
30	3.903	1.646	1.092	1.080	0.012

This table reports the out-of-sample loss/gain induced by following the naïve ($1/N$) investment rule rather than the optimal ex post mean–variance investment rule. Given the out-of-sample standard deviation of the naïve portfolio, the table reports the average return of the *ex ante* (which is not available to investors) and ex post M–V efficient portfolios with the same standard deviation. All numbers are reported as percents. The data consist of monthly returns on the 30 Fama-French value-weighted industry portfolios from January 1996 to May 2007. The in-sample period covers the first half of the sample period, while the out-of-sample period covers the second half

[a]The portfolio is located on the M–V *ex ante* efficient portfolio. Hence, it is not available to investors and given here only for comparison with the other two portfolios

Table 4.5 Out-of-sample analysis of the mean–variance approach – investment constraints

Number of assets (N)	Standard deviation (1)	M–V *ex ante* efficient average return (2)	Ex post M–V average return (3)	Naïve average return (4)	Difference = ex post − naïve [(5) = (3) − (4)]
15	3.912	1.442	0.858	1.071	−0.213
17	3.947	1.606	0.877	1.148	−0.271
20	3.998	1.679	1.220	1.255	−0.035
22	3.998	1.679	1.177	1.185	−0.008
25	4.060	1.700	1.203	1.146	0.057
27	3.997	1.679	1.168	1.111	0.057
30	3.903	1.646	1.144	1.080	0.064

This table reports the out-of-sample loss/gain induced by following the naïve $(1/N)$ investment rule rather than the ex post optimal mean–variance investment rule with constraints on the fraction of wealth allowed to be allocated to each individual asset. Given the out-of-sample standard deviation of the naïve portfolio, the table reports the average return of the *ex ante* and ex post M–V constrained portfolios with the same standard deviation. All numbers are reported as percents. The data consist of monthly returns on the 30 Fama-French value-weighted industry portfolios from January 1996 to May 2007. The in-sample period covers the first half of the sample period, while the out-of-sample period covers the second half

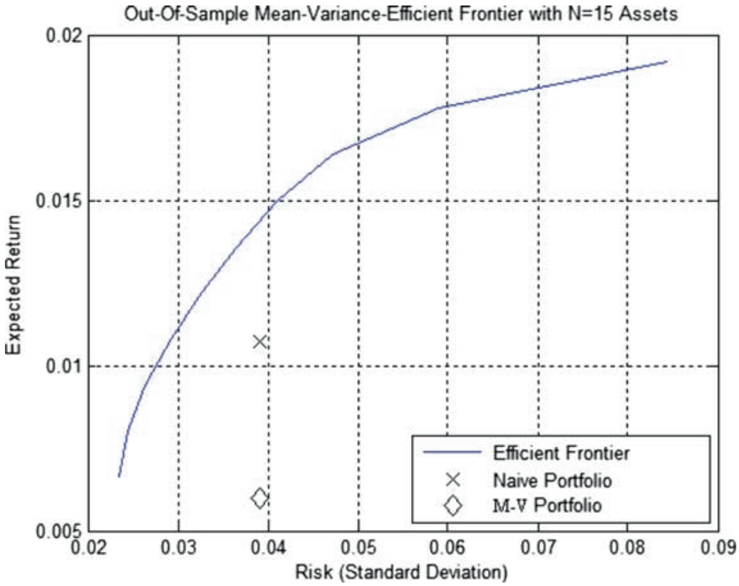


Fig. 4.3 *Out-of-sample* mean–variance analysis of the "$1/N$ rule" when $N = 15$. This figure compares between the *out-of-sample* performance of the "$1/N$ rule" and the Mean–Variance efficient portfolio, relative to the "true" *out-of-sample* mean–variance efficient frontier (which is unavailable), when the number of available assets is 15. Data consist of monthly returns on the 30 Fama-French industry performance from January 1996 to May 2007. The first half of the sample period is optimization period, while the second half is the *out-of-sample* period

these findings visually, plotting the *ex ante* (unavailable) efficient frontier with the M–V portfolio and the "$1/N$ rule" below it. The figures clearly show that the losses are substantial due to the instability of the return distribution, but while the "$1/N$ rule" is clearly superior to the M–V rule when $N = 15$, it is actually inferior to it when $N = 30$.

Can we do any better than the ex post M–V strategy? We try to improve on this strategy by implementing a *constrained* M–V approach that imposes restrictions on the allowable proportions allocated in each asset to mitigate some of the estimation error risk. In particular, we restrict the proportion to no more than 15% when $15 \leq n \leq 20$ and to no more than 10% when $20 \leq n \leq 30$.

Table 4.5 reports the results of the *out-of-sample* analysis with the *constrained* M–V strategy in the same manner they are reported in Table 4.4.[8] For example, the results show that when $N = 15$, the "$1/N$ rule" portfolio outperforms the constrained M–V portfolio by 0.213% (compared to 0.472% with the nonconstrained ex ante M–V strategy). Thus, the results show that in all cases, the *constrained* M–V strategy outperforms the *non*constrained M–V strategy, and when $25 \leq n$, it

[8] Of course one can employ other constraints. Moreover, one can change the constraints until the best *ex ante* results are obtained.

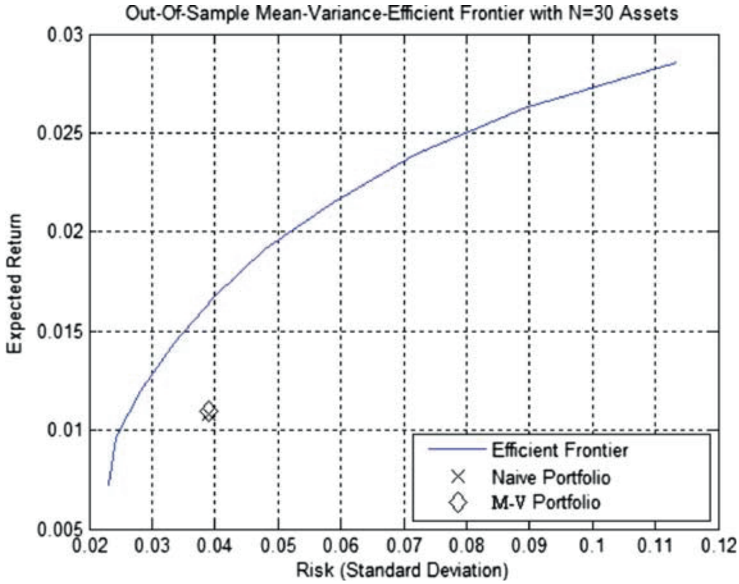


Fig. 4.4 *Out-of-sample* mean–variance analysis of the "$1/N$ rule" when $N = 30$. This figure compares between the *out-of-sample* performance of the "$1/N$ rule" and the Mean–Variance efficient portfolio, relative to the "true" *out-of-sample* mean–variance efficient frontier (which is unavailable), when the number of available assets is 30. Data consist of monthly returns on the 30 Fama-French industry performance from January 1996 to May 2007. The first half of the sample period is optimization period, while the second half is the *out-of-sample* period

outperforms the "$1/N$ rule" as well. Figures 4.5 and 4.6 illustrate these results by showing that while both the M–V and the "$1/N$" portfolios lie below the *ex ante* efficient frontier, they trade places when $N = 15$ and when $N = 25$.

To sum up, the findings in this subsection suggest that while the "1/*N rule*" is by definition suboptimal in *in-sample*, it might actually outperform the ex post M–V efficient investment in *out-of-sample*. This is especially true when the number of assets is relatively small, because when there are many assets involved, it is usually optimal to invest in a small number of them. We also find that a constrained M–V portfolio is generally better in *out-of-sample* than the unconstrained M–V portfolio, because it mitigates some of the estimation risk. However, it is still inferior to the "1/*N rule*" when the number of assets is small.

4.2.3 Expected Utility Results

Section 4.2.2 analyzed the *in-sample* and *out-of-sample* relative performance of the "$1/N$ rule" within the Mean–Variance framework. Next, we analyze its performance within the more general framework of expected utility theory. Even though

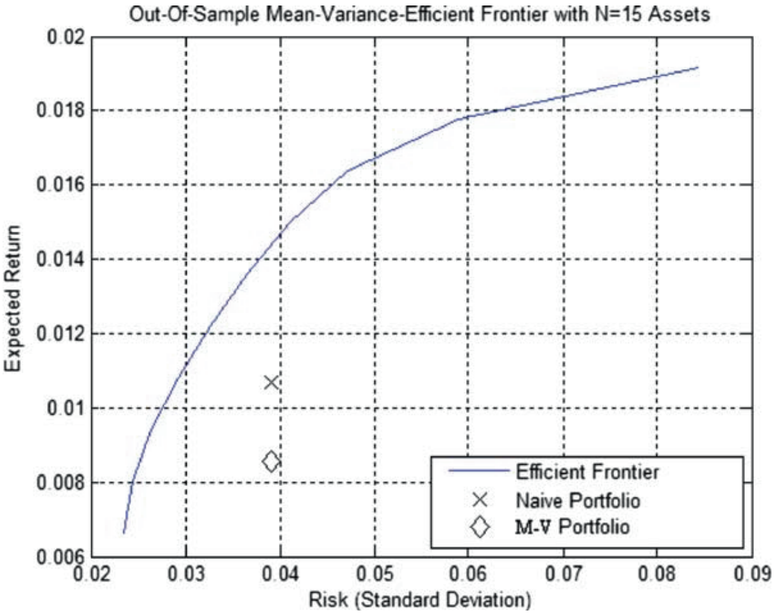


Fig. 4.5 *Constrained out-of-sample* mean–variance analysis of the "$1/N$ rule" when $N = 15$. This figure compares between the *out-of-sample* performance of the "$1/N$ rule" and the constrained Mean–Variance efficient portfolio, relative to the "true" *out-of-sample* mean–variance efficient frontier (which is unavailable), when the number of available assets is 15. Data consist of monthly returns on the 30 Fama-French industry performance from January 1996 to May 2007. The first half of the sample period is optimization period, while the second half is the *out-of-sample* period

we conducted this analysis with a wide range of utility functions and degrees of risk aversion, we concentrate here on the power utility function with the following degrees of relative risk aversion (RRA):

$$U(R) = \frac{(R)^{1-\gamma}}{1-\gamma}, \quad \text{where } \gamma = 2, 4, 6. \tag{4.7}$$

where R is defined as the return on the portfolio investment. Note that the power utility function is characterized by the following desired properties: $U' > 0, U'' < 0, U''' > 0$ [see Arrow (1971)].

We measure the expected utility loss due to the employment of the "$1/N$ rule" for portfolios composed of various number of assets, N, ranging from N=2 to N=30. Given N risky assets and m periods, the empirical parallel of (4.3) becomes:

$$\max_{\underline{x}} \left\{ \frac{1}{m} \sum_{j=1}^{m} U\left(\sum_{i=1}^{n} x_i R_{ij} \right) \right\}, \tag{4.3$'$}$$

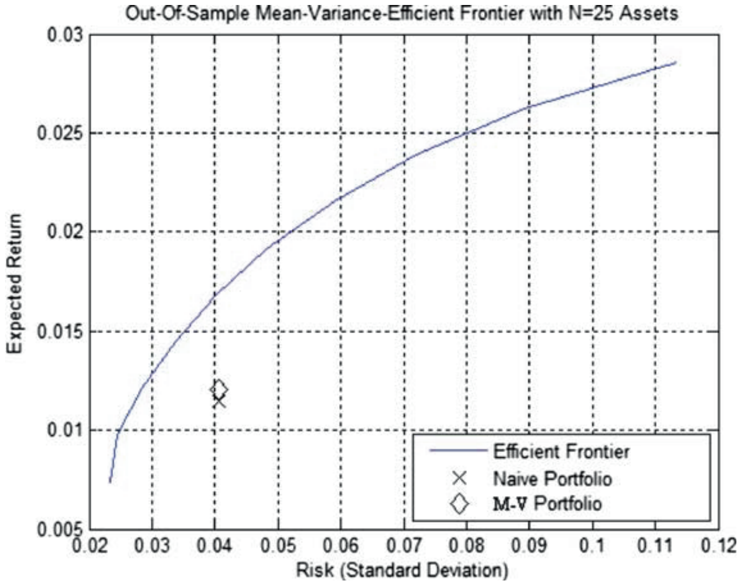


Fig. 4.6 *Constrained out-of-sample* mean–variance analysis of the "$1/N$ rule" when $N = 25$. This figure compares between the *out-of-sample* performance of the "$1/N$ rule" and the constrained Mean–Variance efficient portfolio, relative to the "true" *out-of-sample* mean–variance efficient frontier (which is unavailable), when the number of available assets is 30. Data consist of monthly returns on the 30 Fama-French industry performance from January 1996 to May 2007. The first half of the sample period is optimization period, while the second half is the *out-of-sample* period

where R_{ij} is the return on security i in period j and the rest of the constraints are formulated exactly as in (4.3). Thus, we first calculate the utility of a possible portfolio $\sum_{i=1}^{n} x_i R_{ij}$, then calculate the average utility across all periods. By the maximization method we find the vector $\underline{x}$ which maximizes the expected utility given by ($4.3'$). Thus, we are faced here with a concave target function maximization problem, subject to linear constraints, which we solve using an algorithm from the quasi-Newton family of algorithms.[9]

Table 4.6 reports the *in-sample* results, which reveal that an *in-sample* expected utility loss always exists when one employs the "$1/N$ rule" strategy. The magnitude of this loss varies between \$5.2 and \$28.4 per month for every \$10,000 invested. As before, the loss increases with the number of assets. For example, when the RRA coefficient equals 4, the certainty equivalent loss grows from \$12.3 when $N=2$ to \$25.1 when $N=30$.

[9] We use a Sequential Quadratic Programming routine available from Matlab's optimization toolbox, implemented in the function "fmincon." For further details on the family of concave linearly constrained optimization problems, see Bazaraa and Shetty (1979).

Table 4.6 In-sample analysis of the expected utility approach

Number of assets (N)	Naïve CE	Optimal CE	Difference = optimal − naïve
Panel A: $U(R) = R^{1-\gamma} / \gamma,\ \gamma = 2$			
2	10,059.7	10,068.3	8.6
3	10,081.4	10,090.4	9.0
5	10,087.4	10,099.3	11.9
10	10,082.3	10,099.3	17.0
20	10,099.5	10,118.8	19.3
30	10,090.4	10,118.8	28.4
Panel B: $U(R) = R^{1-\gamma} / \gamma,\ \gamma = 4$			
2	10,040.4	10,052.7	12.3
3	10,059.8	10,065.0	5.2
5	10,072.3	10,079.8	7.5
10	10,068.6	10,079.8	11.2
20	10,081.9	10,098.0	16.1
30	10,072.9	10,098.0	25.1
Panel C: $U(R) = R^{1-\gamma} / \gamma,\ \gamma = 6$			
2	10,020.5	10,027.1	6.6
3	10,037.9	10,046.2	8.3
5	10,056.9	10,063.8	6.9
10	10,054.2	10,065.0	10.8
20	10,063.4	10,077.7	14.3
30	10,054.3	10,077.7	23.4

This table reports the in-sample loss induced by following the naïve $(1/N)$ investment rule rather than the optimal (expected utility maximizing) investment rule. For each number of available assets, the table reports the certainty equivalent (CE) per \$10,000 investment corresponding to the naïve portfolio, as well as the CE corresponding to the expected utility maximizing portfolio. The data consist of monthly returns on the 30 Fama-French value-weighted industry portfolios from January 1996 to May 2007

Furthermore, the results of Table 4.6 show that, generally, utility functions with higher levels of risk aversion exhibit a smaller loss in terms of expected utility due to the employment of the "$1/N$ rule" strategy. This result is due to the fact that optimal portfolios that maximize the expected utility of such utility functions tend to be more diversified, relative to optimal portfolios that maximize utility functions with lower levels of risk aversion. Therefore, the optimal investment proportions corresponding to utility functions with higher levels of risk aversion are closer to the "$1/N$ rule" strategy proportions because they include more assets with nonzero investment proportions, and relatively smaller investment proportions allocated to each asset. Thus, for utility with a high degree of risk aversion the "$1/N$ rule" strategy yields portfolios that are closer to the optimal expected utility maximizing portfolios, and therefore incurs smaller expected utility loss. To illustrate this, note that when $N = 20$, the certainty equivalent loss drops from \$19.3 when RRA $= 2$, to \$16.1 when RRA $= 4$, and to \$14.3 when RRA $= 6$.

How do we interpret these findings? There is a large body of evidence suggesting that households hold a relatively small number of assets in their portfolios and have high RRA coefficients.[10] The above results suggest that even in *in-sample*, when the "$1/N$ rule" is suboptimal by definition, it tends to generate smaller losses when there are fewer assets and when the RRA is higher. Thus, we can conclude that the usage of the "$1/N$ rule" by households is not as irrational as it might seem, because it generates relatively small losses compared to the optimal EU maximizing strategy.[11]

On the other hand, institutional investors hold many assets in their portfolio. As can be seen from Table 4.6, the loss is much larger when $N = 30$. However, it is obvious that institutional investors do not employ the "$1/N$ rule," hence the loss reported in Table 4.6 is irrelevant for institutional investors.

Thus far, we have analyzed the *in-sample* loss due to the "$1/N$ rule" strategy. Nevertheless, as explained above, because the distributions of returns may not be stable over time, the more relevant results for practical investment decision making are the *out-of-sample* rates of return induced by the employment of the "$1/N$ rule" strategy. Tables 4.7 and 4.8 report the results of the *out-of-sample* tests. Table 4.7 does not allow imposing constraints on the proportion allocated to each asset, while Table 4.8 imposes such constraints to mitigate the estimation error risk induced by the EU strategy that is based on past data.

As Table 4.7 reveals, it is no longer the case that the EU strategy is superior. Consistent with our previous findings, the "$1/N$ rule" is the superior strategy when the number of available assets is small and when the degree of relative risk aversion is high. To illustrate, note that when $N = 3$, the "$1/N$ rule" outperforms the EU strategy by \$5.7 when RRA $= 2$, by \$9.1 when RRA $= 4$, and by \$9.8 when RRA $= 6$. However, when the number of available assets is large and the relative risk aversion is low, the EU strategy is better. To see this, note that when $N = 30$, the EU strategy outperforms the "*1/N rule*" by \$27.2 when RRA $= 2$ and by \$20 when RRA $= 4$.

Figures 4.7 and 4.8 depict the Cumulative Distribution Functions (CDF) of the two investment strategies corresponding to N=2 and RRA=6 and to N=30 and RRA=2, respectively. As the figures show, the "*1/N rule*" strategy provides, as expected, an *ex ante* less risky investment. This can be seen due to the heavier tails

[10] Barber and Odean (1999) found that the median household held during the period of 1991–1996 an average of 2.61 individual stocks in their portfolio. In earlier studies, the Federal Reserve Board's survey of Financial Characteristics Consumers found that the average number of securities in a portfolio was 3.41. Moreover, about one-third of individuals with portfolios held only one stock, 50% held two stocks or less, and only 10.7% held more than 10 stocks. [see Blume et al. (1974) and Blume and Friend (1975)]. There is also a large body of evidence regarding the magnitude of the relative risk aversion coefficient. Relying on cross-section data, Friend and Blume (1975) estimate the RRA to be about two. Kandel and Stambaugh (1991) estimate its value to be as large as 29.

[11] Selecting to invest in a very small number of assets does not necessarily contradict expected utility maximization. For example, if one considers fixed and variable transaction costs and when the investment capital is not large, one can rationalize the holding of a very small number of assets. For details, see Levy (1978) and Merton (1987). Yet, in this study we assume that when N assets are held, the "$1/N$ rule" strategy is employed as found by Benartzi and Thaler (2001).

Table 4.7 Out-of-sample analysis of the expected utility approach

Number of assets (N)	Naïve CE	EU CE	Difference = EU − Naïve
Panel A: $U(R) = R^{1-\gamma} / \gamma,\ \gamma = 2$			
2	10,054.3	10,055.5	1.2
3	10,084.0	10,078.3	−5.7
5	10,083.3	10,080.3	−3.0
10	10,085.1	10,092.3	7.2
20	10,109.4	10,119.9	10.5
30	10,092.7	10,119.9	27.2
Panel B: $U(R) = R^{1-\gamma} / \gamma,\ \gamma = 4$			
2	10,045.9	10,037.7	−8.2
3	10,068.9	10,059.8	−9.1
5	10,071.3	10,069.8	−1.5
10	10,074.6	10,080.2	5.6
20	10,092.4	10,096.5	4.1
30	10,076.5	10,096.5	20.0
Panel C: $U(R) = R^{1-\gamma} / \gamma,\ \gamma = 6$			
2	10,037.2	10,028.6	−8.6
3	10,053.4	10,043.6	−9.8
5	10,059.1	10,048.3	−10.8
10	10,063.7	10,069.2	5.5
20	10,074.5	10,083.7	9.2
30	10,059.5	10,083.7	24.2

This table reports the out-of-sample loss/gain induced by following the naïve $(1/N)$ investment rule rather than the optimal (expected-utility maximizing) investment rule. For each number of available assets, the table reports the certainty equivalent (CE) per \$10,000 investment corresponding to the naïve portfolio, as well as the CE corresponding to the *ex ante* expected utility maximizing portfolio. The out-of-sample procedure employs a moving 60-month in-sample window to calculate the optimal portfolio weights, which are then applied to the 61st month to calculate the out-of-sample return. The procedure is repeated until May 2007. The data consist of monthly returns on the 30 Fama-French value-weighted industry portfolios from January 1996 to May 2007

of the CDF corresponding to the expected utility maximizing strategy. However, Fig. 4.7 demonstrates how the difference between the two CDFs decreases for higher values RRA and smaller values of N. While the tails of the expected utility maximizing strategy are much heavier than the tails of the "$1/N$ rule" strategy for RRA=2, the difference is much smaller for RRA=6. The intuitive explanation for this result is that the expected utility maximizing strategy for very high risk-aversion-levels utility functions induced a relatively well-diversified portfolio, and therefore this portfolio is closer to the investment strategy implied by the "$1/N$ rule." Therefore, the two CDF tend to coincide for a relatively high degree of risk aversion.

Let us examine in detail the "$1/N$ rule" in the relevant cases for $N = 2, 3, 4$ and for $\gamma = 4, 6$. In these cases, we find in Table 4.7 that $\text{CE}_{1/N} > \text{CE}_{\max}$ in all these relevant cases, implying a positive gain in expected utility due to the employment of "$1/N$ rule" rather than the max EU$(\cdot)$ rule. Thus, where the most relevant *out-of-sample* returns for households are considered, the experimental and empirical

Table 4.8 Out-of-sample analysis of the expected utility approach: investment constraints

Number of assets (N)	Naïve CE	EU CE	Difference = EU − naïve
Panel A: $U(R) = R^{1-\gamma} / \gamma,\ \gamma = 2$			
15	10,091.7	10,092.0	0.3
20	10,109.4	10,112.1	2.7
25	10,098.0	10,112.1	14.1
30	10,092.7	10,112.1	19.4
Panel B: $U(R) = R^{1-\gamma} / \gamma,\ \gamma = 4$			
15	10,075.5	10,078.6	3.1
20	10,092.4	10,104.7	12.3
25	10,080.6	10,104.7	24.1
30	10,076.5	10,104.7	28.2
Panel C: $U(R) = R^{1-\gamma} / \gamma,\ \gamma = 6$			
15	10,058.7	10,059.3	0.6
20	10,074.5	10,077.6	3.1
25	10,062.3	10,078.8	16.5
30	10,059.5	10,078.8	19.3

This table reports the out-of-sample loss/gain induced by following the naïve ($1/N$) investment rule rather than the optimal (expected-utility maximizing) investment rule. For each number of available assets, the table reports the certainty equivalent (CE) per \$10,000 investment corresponding to the naïve portfolio, as well as the CE corresponding to the *ex ante* expected utility maximizing portfolio. The out-of-sample procedure employs a moving 60-month in-sample window to calculate the optimal portfolio weights, which are then applied to the 61st month to calculate the out-of-sample return. The procedure is repeated until May 2007. The data consist of monthly returns on the 30 Fama-French value-weighted industry portfolios from January 1996 to May 2007

results of B&T and the Talmudic assertion have a very strong foundation within the expected utility framework. The superiority of the "$1/N$ rule" strategy in many of the cases reported above is induced by the instability of the distributions returns and the hedge that the "$1/N$ rule" strategy has against such changes in the distributions over time. Yet, once again, for institutional investors with a relatively large N, the "$1/N$ rule" incurs a large loss (see Table 4.7), but this hypothetical loss is irrelevant as these investors do not employ the naïve rule.

Maximizing expected utility with no investment constraints and the "$1/N$ rule" are two extreme investment strategies. Maximizing expected utility with constraints on the investment proportion provide an investment strategy which falls between the above two extreme strategies. If the constraints are very strong, we are back to the "$1/N$ rule." If the constraints are ineffective we are back to expected utility maximization strategy. Thus, the max EU($\cdot$) strategy with constraints provides a strategy which may enjoy both worlds: on the one hand it is not an arbitrary rule like the "$1/N$ rule," and on the other hand it provides a hedge, albeit a partial one, against changes in the distributions of returns over time. As explained above, we cannot lower the constraints much. For example, for $N = 20$ if the constraint is $p \leq 0.05$, this means that exactly 5% is invested in each asset, and therefore the max EU($\cdot$) policy with these constraints coincides with the "$1/N$ rule" policy. Thus, to have a meaningful difference between the two policies, we need to impose higher constraints than those coinciding with the "$1/N$ rule" strategy.

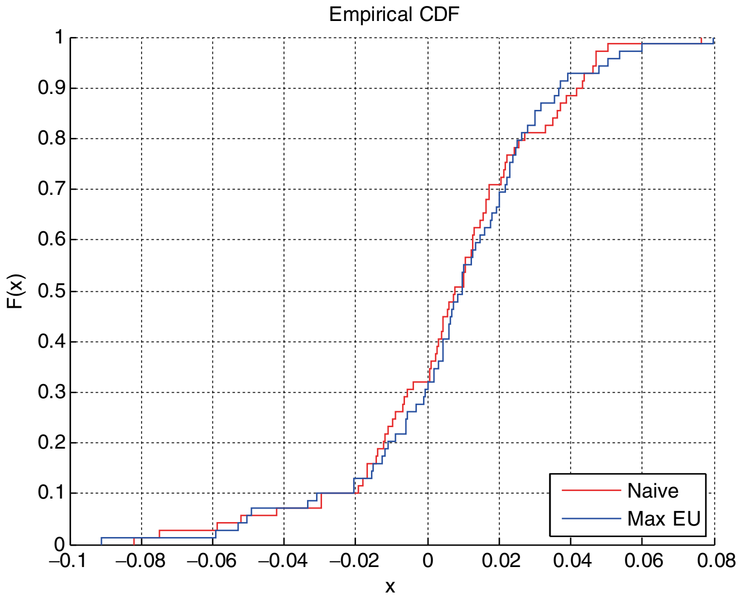


Fig. 4.7 Expected utility analysis of the "$1/N$ rule" when $N = 2$ and RRA $= 6$. This figure compares between the *out-of-sample* performance of the "$1/N$ rule" and the expected utility optimal portfolio, by plotting the cumulative distribution function (CDF) of the *out-of-sample* return vectors corresponding to the two investment strategies, when there are two available assets and investors have a power utility function with a relative risk aversion (RRA) coefficient of 6. Data consist of monthly returns on the 30 Fama-French industry performance from January 1996 to May 2007

Tables 4.8 provides the results for $N \geq 15$. Unlike the previous results, the outcomes here are clear-cut in favor of the max EU($\cdot$) rule with constraints on the investment proportions. In all cases, we have that $CE_{1/N} < CE_{\max}$, i.e., the "1/*N rule*" strategy is inferior to the max EU($\cdot$) strategy with constraints. As the table shows, the difference between the two strategies grows as the number of available assets grows. Yet, it is important to emphasize that Table 4.8 only reports cases with $N \geq 15$. Hence, max EU($\cdot$) with constraints is more appropriate to large investors or institutional investors but not to households, which hold 2–4 assets in their portfolio. Thus, while for household investors, the "$1/N$ rule" performs best in the relevant range of risk aversion, for institutional investors Max EU($\cdot$) with constraint on the investment proportion seems to be the best investment strategy.

4.3 Concluding Remarks

In his breakthrough article, Markowitz developed the Mean–Variance theory which is the foundation of the Sharpe–Lintner CAPM. Using the Mean–Variance theory in practice is difficult because one needs to estimate the various parameters and

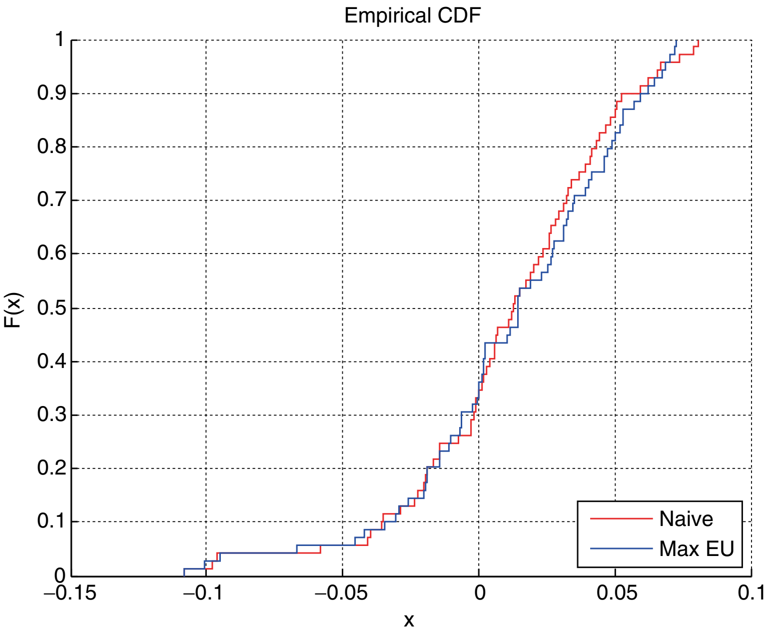


Fig. 4.8 Expected utility analysis of the "$1/N$ rule" when $N = 30$ and RRA $= 2$ This figure compares between the *out-of-sample* performance of the "$1/N$ rule" and the expected utility optimal portfolio, by plotting the CDF of the *out-of-sample* return vectors corresponding to the two investment strategies, when there are two available assets and investors have a power utility function with a RRA coefficient of 6. Data consist of monthly returns on the 30 Fama-French industry performance from January 1996 to May 2007

the investment mix is very sensitive to sampling error in the estimates. The Talmud suggests a naïve diversification strategy. Who is right in practice? Markowitz or the Babylonian Talmud? If the *ex ante* parameters are available, Markowitz is right. If they are unavailable but can be estimated using the ex post parameters, Markowitz is right in some cases (in particular for institutional investors) and the Talmudic wise men have a valid point in others.

Benartzi and Thaler (2001) (B&T) have shown both experimentally and empirically that the 1500-years-old Talmudic recommendation to invest 1/3 in each of the three available assets, i.e., employing the naïve "$1/N$ rule," conforms with what household investors tend to do in practice. Obviously, employing the "$1/N$ rule" ignores the assets' parameters, hence contradicts expected utility theory. In addition, Blume and Friend (1975) and Barber and Odean (1999) show that household investors typically hold 2–4 assets in their portfolio. These two findings contradict expected utility paradigm, in general, and contradict Markowitz's M–V analysis and the Sharpe–Lintner CAPM in particular. Can we say that household investors who employ the "$1/N$ rule" are irrational? Can such a behavior be classified simply as an error that fits into the category of psychological errors, which are well documented in the psychological and experimental economic literature? The answer to this question depends on the type of investors (households or institutional investors) and the stability of the distribution of returns over time.

Our empirical investigation consists of an *in-sample* analysis, where one implicitly assumes that distributions of returns are stable over time, and an *out-of-sample* analysis, which implicitly allows distributions of returns to be unstable over time. In *in-sample*, the "$1/N$ rule" is by definition inferior to optimizing investment strategies such as the M–V or EU strategies. In *out-of-sample*, the "$1/N$ rule" can outperform these strategies, since they rely on past distributions of returns and are therefore exposed to estimation error risk. In fact, the more unstable the empirical distributions are over time, the larger the chance "$1/N$ rule" will prevail, since it contains a hedge against possible errors in the investment policy induced by the fact that the *ex ante* distributions of returns might be much different from the ex post distributions of returns.

We find that the loss induced by the employment of the "$1/N$ rule" is a function of the number of assets held in the portfolio and the degree of risk aversion. The larger the number of assets held and the lower the degree of risk aversion, the larger the loss due to the employment of the "$1/N$ rule." For household investors, who hold a relatively small number of assets in their portfolio and have a relatively high level of risk aversion, we find that the "$1/N$ rule" strategy is superior to expected utility and to M–V maximization in many *out-of-sample* cases. The more unstable the distributions are over time, the larger the advantage of the "1/*N rule*" as it does not suffer from the pitfalls of expected utility maximization, which assign a high investment proportion to an asset with a positive *in-sample* outlier. For institutional investors (i.e., large N), expected utility and Markowitz's portfolio is superior: the "$1/N$ rule" yield inferior *ex ante* performance.

Thus, our relevant *out-of-sample* tests show that for high degrees of risk aversion and for small numbers of assets, which characterize household portfolios in practice, the Talmudic assertion conforms to rational investment policy. For a large number of assets, the Talmudic "$1/N$ rule" suggestion is inferior to Markowitz's M–V portfolio. Of course, one might be able to improve the estimation techniques and show that Markowitz's M–V portfolio is also better for small portfolios. However, this issue is beyond the scope of this paper.

References

Arrow, K.J. (1971), "Essays in the Theory of Risk-Bearing," Makham Publishing Company, Chicago, IL.

Barber, B.M., and Odean, T. (1999), "Trading is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors," *Journal of Finance, 55*, 773–806.

Bazaraa, M., and Shetty, C.M. (1979), "Nonlinear Programming – Theory and Algorithms," Wiley, New York.

Benartzi, S., and Thaler, R.H. (2001), "Naive Diversification Strategies in Defined Contribution Saving Plans," *American Economic Review, 91*, 79–98.

Best, M.J., and Grauer, R.R. (1991), "On the Sensitivity of Mean–Variance–Efficient Portfolios to Changes in Asset Means: Some Analytical and Computational Results," *Review of Financial Studies, 4*, 315–342.

Blume, M.E., and Friend, I. (1975), "The Asset Structure of Individual Portfolios and Some Implications for Utility Functions," *Journal of Finance, 30*, 585–603.

Blume, M.E., Crocket, J., and Friend, I. (1974), "Stock Ownership in the United States: Characteristics and Trends," *Survey of Current Business, 54*, 16–40.

Frost, P.A., and Savarino, J.E. (1986), "An Empirical Bayes Approach to Efficient Portfolio Selection," *Journal of Financial and Quantitative Analysis, 21*, 293–305.

Frost, P.A., and Savarino, J.E. (1988), "For Better Performance: Constrain Portfolio Weights," *Journal of Portfolio Management, 5*, 29–34.

Green, R.C., and Hollifield, B. (1992), "When will mean–variance efficient portfolios be well diversified?," *Journal of Finance, 47*, 1785–1809.

Jagannathan, R., and Ma, T. (2003). "Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps," *Journal of Finance, 58*, 1651–1683.

Kandel, S., and Stambaugh, R. (1991), "Asset Returns and Intertemporal Preferences," *Journal of Monetary Economics, 27*, 39–71.

Kroll, Y., Levy, H., and Markowitz, H.M. (1984). "Mean–Variance Versus Direct Utility Maximization," *Journal of Finance, 39*, 47–61.

Kroll Y., Levy H., and Rapoport A. (1988), "Experimental Tests of the Separation Theorem and the Capital Asset Pricing Model," *American Economic Review, 78*, 500–519.

Levy, H. (1978), "Equilibrium in an Imperfect Market: A Constraint on the Number of Securities in the Portfolio," *American Economic Review, 64*, 643–659.

Levy, H. (1983), "The Capital Asset Pricing Model: Theory and Empiricism," *The Economic Journal, 93*, 145–165.

Levy, H. (2006), "Stochastic Dominance: Investment Decision Making Under Uncertainty," 2nd edition, Springer, New York.

Levy, H., and Markowitz, H.M. (1979), "Approximating Expected Utility by a Function of Mean and Variance," *American Economic Review, 69*, 308–317.

Levy, H., and Sarnat, M. (1972), "Investment and Portfolio Analysis," Wiley Series in Finance, New York.

Lintner, J. (1965), "Security Prices, Risk, and the Maximal Gains from Diversification," *Journal of Finance, 20*, 587–615.

Markowitz, H.M. (1952a), "Portfolio Selection," *Journal of Finance, 7*, 77–91.

Markowitz, H.M. (1952b), "The Utility of Wealth," *Journal of Political Economy, 60*, 151–156.

Markowitz, H.M. (1991), "Foundation of Portfolio Theory," *Journal of Finance, 7*, 469–477.

Merton, R. (1987), "An Equilibrium Market Model with Incomplete Information," *Journal of Finance, 42*, 483–510.

Sharpe, W.F. (1964), "Capital Asset Prices: A Theory of Market Equilibrium," *Journal of Finance, 19*, 425–442.

Von-Neumann, J., and Morgenstern, O. (1953), "Theory of Games and Economic Behavior," 3rd edition, Princeton University Press, Princeton, NJ.

Zweig, J. (1998), "Five Investing Lessons from America's Top Pension Funds," *Money* (January), 115–118.

Chapter 5
On the Himalayan Shoulders of Harry Markowitz

Paul A. Samuelson

Few scientific scholars live to see their brain children come into almost universal usage. Harry Markowitz (1952, 1959, 2008) has been such an exceptional innovator. His quadratic programming Mean-Variance algorithms are used daily by thousands of money managers everywhere.

When a quantum upward jump occurs, Robert K. Merton and other historians of science tell us that usually more than one scholar contributes to the advance – as with Newton and Leibniz or Darwin and Wallace. When we cite Markowitz–Tobin–Lintner–Mossin–Sharpe methodologies, we pay tribute to the creative interactions among and between the innovators.[1]

Genuine scientific advances all too often do meet with resistance from historical orthodoxies. Max Planck (1900, 1901) gained eternal fame for himself when ("as an act of desperation") he introduced *quantum* notions into classical physics. Was he instantly and universally applauded? Not quite so. Autobiographically, he had to declare that old guards are slow to accept new-fangled theories. As they are so often resistant to new methodologies, the new orthodoxy gets born only after they die one by one. Planck sums it up: *Science progresses funeral by funeral*!

Harry Markowitz encountered the Planckian syndrome early on. At Markowitz's Chicago 1952 oral Ph.D. exam, Professor Milton Friedman made waves against quadratic programming, declaring that it was not even economics, and neither was it interesting mathematics.

Any savant can have a bad hair day. But again, almost half a century later, when Markowitz shared with Merton Miller and Bill Sharpe the 1990 Nobel Prize, Dr. Friedman told the Associated Press that those awardees would not be on anyone's long list of 100 deserving names. Was this Chicago jingoism? Hardly. Markowitz and Miller were both Chicago guys. (However, maybe Markowitz owed more to his Rand think-tank sojourn than to his years on the Midway? Bill Sharpe, a West Coast UCLA product, did learn much around Rand from Chicagoan Markowitz.)

P.A. Samuelson (✉)
Massachusetts Institute of Technology, Department of Economics, E52-38C Cambridge, MA 02139

[1] Tom Kuhn had documented in MIT lectures that the basic thermodynamic Law of Conservation of Energy owed much to at least a dozen quasi-independent researchers. And no two of them gave identical interpretations and nuances!

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_5,
© Springer Science+Business Media, LLC 2010

One's ignorance can be self-costly. The oral records in the TIAA-CREF files reveal that Trustee Dr. Friedman believed the heresy that investing for four independent periods necessarily mandated being more heavily in risky stocks than would be the case for two or one period investment. Elementary, my dear Dr. Watson." Elementary, yes. But quite wrong, Dr. Friedman. Samuelson (1969) is one of dozens of references that explicate the very common Friedman fallacy.

It was no accident that Harry's 1959 classic book was published at Yale and not at Chicago. In the mid-fifties, Chicago had spewed out its Cowles team of Koopmans, Marschack, and the other Nobel stars. Strangely, it made no Herculean efforts to keep Kenneth Arrow from going to Stanford.

In sports, records are made to be broken. In cumulative science, each new breakthrough will in time be challenged by new generations. Even peerless Albert Einstein, after repeated triumphs during the first two-thirds of his scientific life, spent the last third on lost causes.

"What have you done for me lately?" is science's ungrateful dictum.

Having been stimulated by Markowitz early and late, I compose this present brief note in homage to Harry. It does stray a bit beyond quadratic programming to general Laplacians' Kuhn-Tucker concave programming. Explained here is the tool of an investment gamble's "Certainty Equivalent," which for Laplacians will be (*necessarily*) an "Associative Mean" of feasible wealth outcomes. This procedure costlessly bypasses people's *subjective* utility function for risk. An *external* jury of observers can hone in exactly on any guinea pig's certainty-equivalent risk functions from that person's decisions among alternative risk options.

I also discuss whether the Modigliani-Miller theorem could validly tell Robinson Crusoe (or any society) that the *degree of leveraging and diversifying can be a matter of indifference*; also I discuss possible exceptions to the Tobin (1980) (one-shoe-fits-all) theorem, which defines *unique* proportions of risky stocks allegedly invariant with respect to even huge differences among different people's degree of risk aversion. Also, the Sharpe (1970) ratio can be shown to become problematic as a normative guide. Finally, the present model is a good one to test what it means to define when a risk-taker is or is not being adequately (optimally) rewarded for bearing risk.

5.1 Crusoe Saga to Test for a Society's Optimal Diversification and Leveraging

Yale's great physicist, Willard Gibbs, once wisely said, "Mathematics is a language." Here for brevity and clarity, I shall mostly use math-speak in place of words.

I posit a single Robinson Crusoe on his isolated island. He lives on corn harvest produced by his unit labor: $L = 1$. (Ignorable land is superabundant and therefore free.) He knows only two alternative technologies, A and B.

A: 1^{A} of $L^{\mathrm{A}} \rightarrow 1^{\mathrm{A}}$ of corn, with certainty in both wet and dry seasons
B: 1^{B} of $L^{\mathrm{B}} \rightarrow \frac{1}{4}\mathrm{B}$ of corn when wet season prevails, or 4^{B} of corn when dry season prevails.

Crusoe knows that wet or dry seasons are random variables with 50–50 probabilities:

$$p_{\text{wet}} = p_{\text{dry}} = \frac{1}{2} = p_j, j = 0, 1. \tag{5.1}$$

I add that Crusoe is a "Laplacian" à la Daniel Bernoulli (1738), whose stochastic choices target *maximal* Expected Utility of Outcomes:

$$\text{Max}\,\Sigma_0^1 p_j U(W_j) \equiv \text{Max}\left\{\frac{1}{2}U(W_0) + \frac{1}{2}U(W_1)\right\} \tag{5.2}$$

where (W_0, W_1) are *dry* corn harvests and *wet* corn harvests, respectively.

The new-ish tool I utilize primarily here is the concept of a gambler's or an investor's stochastic "Certainty Equivalent." For Laplacians, it defines itself as $W_{1/2}$ in the following way:

$$\frac{1}{2}U(W_0) + \frac{1}{2}U(W_1) = U\left(W_{1/2}\right) \tag{5.3}$$

$$W_{1/2} = M[W_0, W_1] = \text{Certainty Equivalent function} \tag{5.4}$$

$$\equiv U^{-1}\left\{\frac{1}{2}U(W_0) + \frac{1}{2}U(W_1)\right\} \tag{5.5}$$

where the inverse function

$$U^{-1}\{y\} = x \longleftrightarrow y = U(x) \text{ notationally.} \tag{5.6}$$

The M[W_0,W_1] function is a *general* mean of corn outcomes. But also *singularly*, it is what mathematicians call an Associative Mean (or quasi-linear mean) à la Hardy–Littlewood–Pólya (1952). Equations (5.7) and (5.8) describe M's full content:

$$\text{Min}\{W_0, W_1\} \leq M[W_0 W_1] \leq \text{Max}[W_0, W_1] \tag{5.7}$$

$$M\,[W, W] \equiv W. \tag{5.8}$$

Without proof, I give one of many Abel-like functional equations that our Laplacian Associative Mean must obey, by definition:

$$M[W_0, W_1] \equiv W_{1/2}; M\left[W_0, W_{1/2}\right] \equiv W_{1/4}, M\left[W_{1/2}, W_1\right] \equiv W_{3/4}. \tag{5.9}$$

Then

$$M\left[W_{1/4}, W_{3/4}\right] \text{ must exactly equal } M\,[W_0, W_1]! \tag{5.10}$$

In short,

$$M\,[W_0, W_1] \equiv M\,(M\,\{W_0, M\,[W_0, W_1]\}\,,\ M\,\{W_1, M\,[W_0, W_1]\})\,. \tag{5.11}$$

This math-speak esoteric stems only from the following trivial arithmetic tautology:

$$U\left(W_{1/2}\right) = \frac{1}{2}U(W_0) + \frac{1}{2}U(W_1) \tag{5.12}$$

$$\begin{aligned} U\left(W_{1/4}\right) &= \tfrac{1}{2}U(W_0) + \tfrac{1}{2}\left\{\tfrac{1}{2}U(W_0) + \tfrac{1}{2}U(W_1)\right\} \\ &= \left(\tfrac{3}{4}\right)U(W_0) + \tfrac{1}{4}U(W_1) \end{aligned} \tag{5.13}$$

$$\begin{aligned} U\left(W_{3/4}\right) &= \tfrac{1}{2}U(W_1) + \tfrac{1}{2}\left\{\tfrac{1}{2}U(W_0) + \tfrac{1}{2}U(W_1)\right\} \\ &= \tfrac{3}{4}U(W_1) + \tfrac{1}{4}U(W_0) \end{aligned} \tag{5.14}$$

$$\therefore\ \frac{1}{2}U\left(W_{1/4}\right) + \frac{1}{2}U\left(W_{3/4}\right) = \left(\frac{3}{8} + \frac{1}{8}\right)U(W_0) + \left(\frac{1}{8} + \frac{3}{8}\right)U(W_1) \tag{5.15}$$

$$= \frac{1}{2}U(W_0) + \frac{1}{2}U(W_1).\ \text{QED} \tag{5.16}$$

5.2 Modigliani-Miller Misunderstood

Exam takers too often write: It is a matter of indifference whether General Motors or General Electric increases each shareholder's leverage by floating debt a little or a lot. Whatever the firm spikes up (or down) can be offset by each shareholder's counter algebraic leveraging. What holds for any one corporation allegedly holds (sic) for society. A specified society can allegedly and indifferently leverage (and diversify) little or much.

Let us test this misunderstanding using these above equations from (5.1) to (5.16) for a special Crusoe, who happens to ape the medium risk tolerance of 1738 Daniel Bernoulli. Bernoulli happened to believe that when wealth doubles its *marginal* utility – $dU(W)/dW$ – halves. Such a $U(W)$ must be

$$U(W) = \log W;\ \frac{1}{2}U(W_0) + \frac{1}{2}U(W_1) \equiv \frac{1}{2}\log W_0 + \frac{1}{2}\log W_1. \tag{5.17}$$

Skipping mathematical proof, I will assert that this Crusoe – call him Dan Crusoe – will necessarily be a "Geometric Mean maximizer," i.e., for him

$$M\,[W_0, W_1] \equiv \text{GM} \equiv \sqrt{W_0 \cdot W_1}. \tag{5.18}$$

What will then have to be his stochastic choice between applying his $L = 1$ to A only, to B only, or to some combination of positive fractional L^A and L^B? Applied to Equation (5.18) the A and B parameters, we calculate:

$$\sqrt{\frac{1}{4}^{B} \cdot 4^{B}} = 1^{B} = \sqrt{1^{A} \cdot 1^{A}} = 1^{A}. \tag{5.19}$$

This demonstrates that A only and B only are indifferently choosable. His GM function can also tell the story that, as if by magic, diversification between both A and B guarantees a Certainty Equivalent corn harvest *above* (!) 1^A or 1^B.

By repeated trial and error, or by use of baby Newtonian calculus, Dan can discover that his *maximal* Certainty Equivalent will be 25% above 1, achievable only when he divides his labor (singularly!) 50–50.

For $L^A = \frac{1}{2} = L^B$, the GM becomes

$$\sqrt{\left[\frac{1}{2}(1^{A}) + \frac{1}{2}\left(\frac{1}{4}^{B}\right)\right] \times \left[\frac{1}{2}(1^{A}) + \frac{1}{2}(4^{B})\right]} = \sqrt{\frac{5}{8} \cdot \frac{5}{2}} \tag{5.20}$$

$$= \left(\frac{5}{4}^{A\&B}\right)^{*} > (1^{A})^{*} = (1^{B})^{*}. \tag{5.21}$$

The exact calculus algorithm to find Crusoe's optimal fractional value for L^B–call it x^*–is as follows:

$\underset{x}{\mathrm{Max}}(\mathrm{GM})^2 = \mathrm{Max}\, f(x)$, where

$$U(x) = \left(1 - \frac{3}{4}x\right)(1 + 3x) \tag{5.22}$$

$$U'(x) = \left(2\frac{1}{4}\right) - 2\left(2\frac{1}{4}\right)x \tag{5.23}$$

$$U'(x^*) = \left(2\frac{1}{4}\right) - 2\left(2\frac{1}{4}\right)x^* = 0 \tag{5.24}$$

$$\therefore x^* = \frac{1}{2}^{*} \text{QED.} \tag{5.25}$$

For $\left(\frac{1}{16}, 16\right)$ or (N^{-1}, N), the same singular $\frac{1}{2}^{*}$ must occur for Crusoe with log utility.

Equations (5.17)–(5.25) justify the paradox that embracing extra volatility *cum* a cash quencher can jack up a risk-taker's Certainty Equivalent through the magic of one's optimal diversifying. Technology B, when half its volatility is quenched by safe A, increases Crusoe's Certainty Equivalent by 25%, because $1 + \frac{1}{4}$ is 25% greater than 1 + 0. Equations (5.17)–(5.25) spelled out that story.

Energetic readers can test their comprehension by replacing the A and B story by an A and C story, where C has replaced $\left(\frac{1}{4}^{\mathrm{B}}, 4^{\mathrm{B}}\right)$ by $\left(\frac{1}{16}^{\mathrm{C}}, 16^{\mathrm{C}}\right)$. C's extreme volatility *cum* $L^A = \frac{1}{2} = L^C$, achieves a certainty equivalent far better than 25%, as readers can verify.

5.3 Summary

Societies that differ in individuals' risk tolerance will, and should, end up differing much in their equilibrium degree of diversification. A thousand Crusoes just like Dan, on similar islands and with similar known technologies and similar endowments of productive inputs, will all cleave to the same well-defined, one-and-only optimal degree of diversification.

5.4 Mutual Reinsurances' Twixt Different-Type Persons

By contrast, if say 500 islands are populated by Harriet Crusoe Harmonic Mean maximizers and say 500 are populated by Dan Crusoe Geometric Mean maximizers, then much as in the 1817–1848 Ricardo-J.S. Mill theory of trade according to comparative advantage, both groups can be made better off by the existence of folks *unlike* themselves. On another occasion, I hope to show how the Dan Crusoes will trade (in complete Arrow-Debreu markets) with the more risk-averse Harriet Crusoes. Guess whether the Harriets will import or export corn in dry seasons with the Dan's and whether in wet seasons the direction of trade gets qualitatively reversed.

People who differ can thus mutually benefit from insurance contracts with each other. However, I warn that such Pareto-optimal equilibria need not be at all "fair." Notice that for ethicists who value equally Harriet and Dan types, an exogenous increase in relative number of Harriets "unfairly" impoverishes Harriets and fructifies Dans. (When Genghis Khan gives himself *all* which is produced, that is a Pareto Optimal state in the sense that every departure from it hurts someone. However, what is just or fair or ethical when that happens?)

5.5 A Few Final Words on Quadratic Utility

A risk-averse Crusoe could have had, for certain ranges of corn outcomes, a quadratic utility of the form:

$$U(W) = W - \frac{1}{2}bW^2;\ b > 0, W < b^{-1}, \tag{5.26}$$

where large positive b parameter implies greater risk aversion. Thus a Laplacian Dan Crusoe and a Laplacian Harriet Crusoe could differ in that Dan's b would be less than Harriet's b. Both of them would be queer birds, who become more risk averse at large wealth than at small wealth.

Markowitz–Tobin–Sharpe quadratic programming could handle such birds. And for them the Sharpe ratios and Tobin one-shoe-fits-all riskiness might well apply.

Moreover, my special wet vs. dry universal effects that are common to everybody has permitted Harvard Ph.D. Erkko Etula to write out for publication the Markowitz quadratic programming story in terms of means and variances (μ, σ^2) that I have largely ignored. However, that precise story would be a sub-optimal usage for (μ, σ^2) in the case of the Laplacians and I have described who are Geometric Mean and Harmonic Mean maximizers.

Mr. Etula has shown how and when the Tobin theorem does or does not apply. And Etula's equations for Certainty Equivalents, expressed in terms of μ and σ, can replace the equation(s) for quadratic utility. By contrast, for log W utility, $\sqrt{W}$ utility and -1/W utility respectively:

$$\text{For GM, } M\,[W_0, W_1] \equiv M[\mu - \sigma, \mu + \sigma] \equiv \sqrt{(\mu - \sigma)\,(\mu + \sigma)} \equiv \sqrt{\mu - \sigma^2} \tag{5.27}$$

$$\text{For HM, } M\,[W_0, W_1] \equiv [(\mu - \sigma)(\mu + \sigma)]/\mu = \mu - \sigma^2\mu^{-1} \tag{5.28}$$

$$\begin{aligned}\text{For } M\,[W_0, W_1] &= \left[\frac{1}{2}\sqrt{W_0} + \frac{1}{2}\sqrt{W_1}\right]^2 \\ &= M[\mu - \sigma, \mu + \sigma] = \left[\frac{1}{2}\sqrt{\mu - \sigma} + \frac{1}{2}\sqrt{\mu + \sigma}\right]^2\end{aligned} \tag{5.29}$$

$$\text{For } U(W) \ = W - \frac{1}{2}bW^2, W < \mathrm{b}^{-1},$$

$$M\,[W_0, W_1] = \mu - \frac{1}{2}b(\mu^2 + \sigma^2), \mu + \sigma < b^{-1} \tag{5.30}$$

Notice that in every one of these Laplacian risk-averse cases, higher μ is good and *ceteris paribus* higher σ^2 is bad. Sharpe could find four different "Sharpe ratios" to recommend.

As Etula has confirmed, when wet-dry dichotomy is replaced by [very wet, little wet, no wet at all,...], μ and σ's first two moments cannot tell the whole important skewness, kurtosis, and other vital parts of the story. This has been insufficiently recognized for decades. Only for Itô-Merton–Bachelier *instantaneous* probabilities is quadratic programming precisely accurate. (See Samuelson, 1970.)

5.6 Settling What Must be the Optimal Return for the Risks One Takes

The simple, clean Crusoe case can explicate just what has to be the reward a Laplacian needs to get to compensate for the riskiness chosen to be borne.

By the Newtonian choosing of Dan Crusoe's best Geometric Mean, he has stopped taking on extra riskiness just where his Certainty Equivalent drops off net from its maximum. Recall Equations from (5.20) to (5.25).

Therefore, Dan Crusoe and Harriet Crusoe, or Markowitz–Tobin–Sharpe quadratic programmers, do end up with optimized Equation (5.30)'s $\mu - b(\mu^2 + \sigma^2)$ at the point where marginal riskiness is in definitive balance with marginal reward.

No rational Laplacian fails to be rewarded for the riskiness they choose to bear in the singular $U = W_{-1/b}W^2$ case QED.

Acknowledgments I acknowledge the help from Erkko Etula, patient editorial help from Janice Murray, and from the Alfred P. Sloan Foundation for my MIT post-doctoral fellowship.

References

Bernoulli, D. 1738. Exposition of a new theory on the measurement of risk, *Econometrica* 22 (January 1954): 23–36, a translation from the Latin.

Hardy, G.H., Littlewood J.E., and Pólya G.. 1952. Inequalities. Cambridge, UK and New York, Cambridge University Press.

Markowitz, H. 1952. Portfolio selection, *J Finance* 7. 1:77–91.

Markowitz, H. 1959. Portfolio Selection, Efficient Diversification of Investment. Monograph 16. Cowles Foundation for Research in Economics at Yale University, New Haven, Connecticut.

Markowitz, H. 2008. CAPM investors do not get paid for bearing risk: A linear relation does not imply payment for risk, *J Portfolio Manag* 34 2:91–94.

Planck, M. 1900, 1901. On the law of distribution in energy in the normal spectrum, *Ann Phys* 4:553ff.

Samuelson, P.A. 1969. Lifetime Portfolio Selection by Dynamic Stochastic Programming, *Rev Econ Statistics* 51. 3:239–246.

Samuelson, P.A. 1970. The fundamental approximation theorem of portfolio analysis in terms of means, variances, and higher moments, *Rev Econ Statistics* 37. 4:537–542.

Sharpe, W. 1970. Portfolio theory and capital markets. New York, NY, McGraw-Hill Book Company.

Tobin, J. 1980. Asset accumulation and economic activity. Oxford, Basil Blackwell.

Chapter 6
Models for Portfolio Revision with Transaction Costs in the Mean–Variance Framework

Andrew H. Chen, Frank J. Fabozzi, and Dashan Huang

6.1 Introduction

In 1952, Harry Markowitz formulated a framework for portfolio selection based on two parameters: mean and variance. Over the next five decades, the basic portfolio selection model he formulated has been applied to areas such as capital budgeting and risk management (see Fabozzi et al. 2002) for a further discussion of these applications). Although financial economists immediately recognized the benefit of Markowitz's portfolio selection model, acceptance by asset managers was slow (Rosenberg 1974; Markowitz 1976; Rudd and Rosenberg 1979, 1980; Fama and French 1992, 1995; Bloch et al. 1993; Grinold and Kahn 1999; Chan et al. 1996, 1999; Markowitz 2000; Jacobs et al. 2005). Of course, in the early years, acceptance was stymied by inadequate computer power needed to solve for the efficient portfolios for even a small universe of candidate assets. As a result, early research focused on more computationally efficient models for solving for the optimal portfolio. Other implementation-oriented research addressed the estimation errors associated with the parameters and their impact on the optimal portfolio. For example, Black and Litterman (1992), as well as others (Best and Grauer 1991; Chopra 1993; Chopra and Ziemba 1993), have pointed out that for the mean–variance framework, a small perturbation of the inputs may result in a large change in the optimal portfolio, implying that the parameters should be estimated precisely enough, especially the expected return. Challenges to the assumption that the standard deviation is the appropriate measure of risk in light of the preponderance of empirical evidence that suggests that the return distribution for major asset classes is not normally distributed have lead to alternative formulations of the portfolio selection model (see,

A.H. Chen
Edwin L. Cox School of Business, Southern Methodist University, Dallas, TX, USA

F.J. Fabozzi (✉)
Yale School of Management, New Haven, CT, USA
e-mail: frank.fabozzi@yale.edu

D. Huang
Olin School of Business, Washington University, St. Louis, MO, USA

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_6,
© Springer Science+Business Media, LLC 2010

e.g., Rachev et al. 2005). The first extension of the single-period Markowitz model to a multiperiod setting was formulated by Mossin (1968) using a dynamic programming approach. The model called for the application of the Markowitz single-period model at the beginning of each period.

A very practical issue in the portfolio selection process that is ignored in the mean–variance framework is the treatment of transaction costs. In the application of the mean–variance model, an asset manager only starts with cash and then constructs the optimal portfolio when the portfolio is initially constructed from proceeds received from a client or, in the case of a mutual fund, when proceeds are obtained from the sale of fund shares. What is more common in practice is that an asset manager has an existing portfolio of assets and with the passage of time or changes in expectations regarding the model's inputs will rebalance or revise the portfolio. Revising a portfolio requires taking into consideration the transaction costs or transaction costs associated with the liquidation of positions and the acquisition of new positions (Sharpe 1971, 1984, 1987). This is also true in transition management (i.e., where a client removes a portfolio from one asset management firm and engages another firm to manage those assets). In the restructuring of the portfolio, recognition must be given to transaction costs. Neither the single-period Markowitz model nor its extension to the multiperiod case take into consideration transaction costs in the construction of the optimal portfolio.

In this chapter, we revisit the portfolio revision problem with transaction costs, exploring the impact of transaction costs incurred in portfolio revision on the optimal portfolio position in the traditional mean–variance framework. We focus on static portfolio revision (single period), that is, on revision based on a single decision at the start of the horizon, although the revision may cover several time periods (multiple periods). The main reason for limiting analysis is that current multiperiod portfolio strategies, such as multistage mean–variance criteria, are employed in a myopic manner and the decision maker in each period maximizes the next-period objective. Although this strategy allows analytical tractability and abstracts from dynamic revising considerations, there is growing evidence that the temporal portfolio revision may compromise a significant part of the total risky asset demand (Campbell and Viceira 1999; Brandt 1999). Actually, the key issue has been the inability to directly apply the "principle of optimality" of traditional dynamic programming approach due to the failure of the iterated expectations property for the mean–variance objectives.

Our contribution in this chapter is twofold: (1) we consider the portfolio revision problem with transaction costs that are paid at the end of the planning horizon, and present some analytical solutions for some special cases in the mean–variance framework; and (2) we perform a simple empirical experiment actual market data to show that the impact of the transaction costs is significant, confirming the findings of Chen et al. (1971) that transaction costs should be integrated into the portfolio revision optimization problem, and that lower revision frequency may reduce the magnitude of the impact.

The rest of this chapter proceeds as follows. A review of the literature is provided in Sect. 6.2. Section 6.3 describes the basic formulation for the portfolio revision problem with transaction costs. Section 6.4 considers the portfolio revision decision of Chen et al. (1971) in the mean–variance framework. The estimation error for the inputs, particularly the expected return, usually entails extreme positions in the assets of the optimal portfolio and delivers a poor out-of-sample performance. To reduce the uncertainty in the expected return, in Sect. 6.4.1 we investigate the effect of transaction costs with the model recently proposed by Zhu et al. (2007). Section 6.4.2 gives a simple practical application, followed by a summary of our conclusions in Sect. 6.5.

6.2 Review of Literature

Fifteen years after Markowitz's model was published, Smith (1967) proposed a model for managing a multiperiod portfolio. Rather than the complete revision of the portfolio when expectations of a portfolio manager change making the existing portfolio suboptimal, Smith proposes a heuristic approach that he calls "controlled transition" that permits suboptimal portfolio on a continuing basis but which result in fewer portfolio revisions.

The issue of whether the periodic revision of an efficient portfolio improves performance relative to a strategy of buying and holding an initially efficient portfolio was studied by Johnson and Shannon (1974). Their empirical findings suggest that quarterly revision does in fact far outperform a policy of buying and holding a portfolio that was initially efficient. However, their results were questioned by Bloomberg et al. (1977).

The development of more sophisticated modeling of portfolio revision with transaction costs was first provided by Chen et al. (1971). They propose a model of portfolio revision with transaction costs for both the Markowitz single-period and Mossin multiperiod dynamic programming model. Their model assumes that transaction costs are a constant proportion of the transaction's dollar value. They also suggest how to solve the model when there are multiple periods and multiple assets.

Subsequent to the formulation of the Chen-Jen-Zionst model, there were several important contributions in the past two decades. Davis and Norman (1990) study the consumption and investment decision in the case where the optimal buying and selling policies are charged equal transaction costs. Adcock and Meade (1994) add a linear term for the transaction costs to the mean–variance risk term and minimize this quantity. Yoshimoto (1996) proposes a nonlinear programming algorithm to solve the problem of portfolio revision with transaction costs. Konno and Wijayanayake (2001) consider a cost structure that is concave in the mean absolute deviation model. Best and Hlouskova (2005) give an efficient solution procedure for the case of proportional transaction costs where the transaction costs are paid at the end of the period, impacting both the risk and the return. Finally, Lobo et al. (2007) provide a model for the case of linear and fixed transaction costs that maximizes the expected return subject to the variance constraint.

Michaud (1998) finds two interesting phenomena: (1) the mean–variance model usually generates high portfolio turnover rates, and (2) the optimal portfolios are often counter-intuitive because they exhibit lack of diversification with extreme allocations into just a few assets. For these reasons, it seems interesting to examine how transaction costs in the mean-risk model affect portfolio turnover and lack of diversification. That is, do transaction costs matter in portfolio revision, and could transfer cost control improve portfolio performance?

6.3 Problem Formulation

In this section, we formulate the portfolio revision problem with transaction costs as a standard mathematical optimization problem on the basis of mean-risk framework. We will employ the following notations

n number of risky assets
e vector with all entries equal to ones
y^0 amount invested in risk-free asset before revision
y amount invested in risk-free asset after revision
x^0 portfolio invested in risky assets before revision
x^{b} buying portfolio on the risky assets
x^{s} selling portfolio on the risky assets
x portfolio invested in risky assets after revision, $x = x^0 + x^{\mathrm{b}} - x^{\mathrm{s}}$
$c^{\mathrm{b}}(\cdot)$ buying transfer cost of risky assets with $c^{\mathrm{b}}(\cdot) \geq 0$ (for cash we set $c_{\mathrm{cash}} = 0$, that is, one only pays for buying and selling the assets and not for moving the cash in and out of the account)
$c^{\mathrm{s}}(\cdot)$: selling transfer cost of risky assets with $c^{\mathrm{s}}(\cdot) \geq 0$;
X: convex set of feasible portfolios;
r_{f}: risk-free rate of interest;
r: actual return of risky assets with expectation $\bar{r} = \mathbb{E}(r)$ and $\bar{r} \neq r_{\mathrm{f}} e$.
Σ ($\succ 0$): covariance matrix of the portfolio return.

Although the concept of the desirability of diversification can be traced back to Daniel Bernoulli who argues by example that it is advisable for a risk-averse investor to divide goods which are exposed to some small danger into several portions rather than to risk them all together, Markowitz is the first one to mathematically formulate the idea of diversification of investment by defining the variance as a measure of economic risk (Markowitz 1999; Rubinstein 2002). Through diversification, risk can be reduced (but not necessarily eliminated) without changing expected portfolio return. Markowitz rejects the traditional hypothesis that an investor should simply maximize expected returns. Instead, he suggests that an investor should maximize expected portfolio return while minimizing portfolio risk of return, implying a trade-off between expected return and risk.

Although the mean-risk model is first proposed for the portfolio optimization problem where the economic conditions and investment opportunities are assumed to be static over the planned horizon, the composition of a portfolio of risky assets, however, generally will change over time because of random outcomes of the returns on its constituent assets in the subperiods prior to the horizon. Adjustment of the proportions of the assets may thus be necessary to reestablish an efficient portfolio at the beginning of each subperiod in the planned interval. The investor would also want to adjust the portfolio composition if his expectations or risk aversion changed. The opportunities to adjust enable the investor to increase his expected utility at the horizon and therefore should be taken into account in making investment decisions. Given that the investment decisions are usually made starting with a portfolio of assets rather than cash, some assets must be liquidated to permit investment in other assets, incurring transaction costs in the process of portfolio revision.

More specifically, consider an investment portfolio that consists of holdings in some or all of n risky assets and one risk-free asset. Compared with the traditional portfolio selection on the basis of mean-risk tradeoff, transaction costs in the portfolio revision process add an additional objective to the investor that is competitive to the mean and risk. Suppose the expected return, risk, and transaction costs are $r(x, y)$, $\rho(x)$, and $c(x)$, respectively. In this case, an efficient portfolio must be Pareto efficient. That is, a feasible portfolio $x^* \in X$ is Pareto efficient if and only if there does not exist a feasible $x \in X$ ($x \neq x^*$) such that $r_{x^*} \leq r_x$, $\rho(x^*) \geq \rho(x)$ and $c(x^*) \geq c(x)$ with at least one strict inequality.

Mathematically, a Pareto efficient portfolio solves the following optimization problem:

$$(P_0) \qquad \min_{x \in X} \left\{ -r(x, y), \rho(x), c(x) \right\} \tag{6.1}$$

$$\text{s.t.} \quad x = x^0 + x^{\mathrm{b}} - x^{\mathrm{s}}, \tag{6.2}$$

$$y + e^\top x + c(x) = 1, \tag{6.3}$$

$$x^{\mathrm{b}} \cdot x^{\mathrm{s}} = 0, \tag{6.4}$$

$$x^{\mathrm{b}} \geq 0, \; x^{\mathrm{s}} \geq 0, \tag{6.5}$$

where $x^{\mathrm{b}} \cdot x^{\mathrm{s}} = 0$ is a complementary constraint, i.e., $x_j^{\mathrm{b}} x_j^{\mathrm{s}} = 0$ for $j = 1, \cdots, n$.

In the objective function (P_0), there are three competing objectives: maximizing the portfolio expected return $r(x, y)(= r_{\mathrm{f}} y + \bar{r}^\top x)$ while minimizing both the portfolio risk $\rho(x)$ and the portfolio transaction costs $c(x)$. Notice here that we do not include the position y in $\rho(x)$ and $c(x)$ since we have assumed that increasing or decreasing the risk-free asset does not incur any risk and transaction costs. To find a Pareto portfolio we usually solve the optimization problem by minimizing a tradeoff between the three-objective objectives. That is, (6.1) is equivalent to

$$\min \left\{ -r(x, y) + \gamma \rho(x) + \lambda c(x) \; : \; x \in X \right\},$$

where $\gamma\rho(x)$ and $\lambda c(x)$ are, respectively, the risk and transaction costs the investor will take on, and $\gamma > 0$ is the coefficient of risk aversion (the larger the value of γ, the more reluctant the investor is to take on risk in exchange for expected return). This feature also applies to the transfer cost coefficient λ.

The number x in (6.2) represents the portfolio position to be chosen explicitly through sales x^{s} and purchase x^{b} that are adjustments to the initial holding $x^{\mathbf{0}}$. The second constraint (6.3) is the budget constraint. In contrast to the traditional constraint, there is a new term, the transaction costs. Without loss of generality, we normalize the investor's initial wealth, i.e., $e^{\top}x^{\mathbf{0}} = 1$. The complementary constraint (6.4) and the nonnegative constraint (6.5) rule out the possibility of simultaneous purchases and sales. In practice, simultaneously buying and selling (choosing $x^{\mathrm{b}} > 0$ and $x^{\mathrm{s}} > 0$) can never be optimal because making the allocation to one asset increases and decreases at the same time will give rise to unnecessary transaction costs (see Dybvig (2005) for a detailed discussion). Finally, we generally assume the feasible portfolio set X is such that $X = \{x : Ax \leq 0\}$ where A is an $m \times n$ matrix.

While we leave the specification of portfolio risk $\rho(x)$ until later, we assume the transaction costs to be separable throughout this chapter, that is

$$c^{\mathrm{b}}(x) = \sum_{j=1}^{n} c_j^{\mathrm{b}}(x_j),$$

$$c^{\mathrm{s}}(x) = \sum_{j=1}^{n} c_j^{\mathrm{s}}(x_j),$$

where $c_j(\cdot)$ is the transaction cost function for asset j. We will focus on the proportional transaction cost, proportional to the total dollar value of the selling/buying assets, and investigate its impact on the portfolio revision policy. For the treatment of nonconvex transaction costs, see Konno and Wijayanayake (2001) and Best and Hlouskova (2005) for details. Hence, a balance constraint that maintains the self-refinancing strategy including transaction costs is given as

$$y + e^{\top}x + \sum_{j=1}^{n} c_j^{\mathrm{b}} x_j^{\mathrm{b}} + \sum_{j=1}^{n} c_j^{\mathrm{s}} x_j^{\mathrm{s}} \leq e^{\top}x^{\mathbf{0}} + y^{\mathbf{0}} = 1.$$

In summary, we can equivalently rewrite (P_0) as

$$(P_1) \qquad \begin{aligned} \min_{x\in X} \quad & -(r_{\mathrm{f}}y + \bar{r}^{\top}x) + \gamma\rho(x) + \lambda c(x), \\ \text{s.t.} \quad & x = x^{\mathbf{0}} + x^{\mathrm{b}} - x^{\mathrm{s}}, \\ & y + e^{\top}x + c(x) = 1, \\ & x^{\mathrm{b}} \cdot x^{\mathrm{s}} = 0, \\ & x^{\mathrm{b}} \geq 0, \; x^{\mathrm{s}} \geq 0. \end{aligned} \tag{6.6}$$

We remark that (P_1) usually has some equivalent counterparts in the sense of expected return-risk-cost efficient frontier: (1) maximizing expected return subject to given risk and transaction costs or (2) minimizing risk subject to a given expected return or transaction costs. Although maximizing expected return subject to a given risk and transaction costs may be especially attractive to practitioners who have trouble quantifying their preferences but may have an idea how much volatility and transaction costs are acceptable, we will only focus on problem (P_1) and analyze the efficient frontier by varying the risk-aversion parameter γ and the cost aversion parameter λ. In the following, we present some properties of (P_1) and give some tractable relaxations.

For computational convenience, some papers directly discard the complementary condition (6.6) (see Krokhmal et al. (2001) and Lobo et al. (2007) for instance). Theoretically, discarding the complementary condition may lead to buying and selling a particular asset simultaneously in some cases, which is obviously unreasonable. Actually, if the risk-aversion coefficient γ is very large and the cost-aversion coefficient λ is very small, then we may always select an optimal portfolio with zero weights on the risky asset ($x = 0$) and hence zero risk ($\rho(0) = 0$) by buying and selling the assets or investing all wealth in the risk-free asset. In practice, however, Mitchell and Braun (2004) prove that the intractable complementary constraint (6.6) can be removed in the presence of a risk-free asset. Moreover, the term of transaction costs in the objective function can be replaced by $1 - (y + e^\top x)$, according to the budget constraint, reducing the complexity of (P_1) since $c(x)$ may be complicated such as concave. Then (P_1) can be reduced to the following optimization problem:

$$
(P_2) \qquad \begin{aligned}
\min_{x\in X} \;& -(r_f y + \bar{r}^\top x) + \gamma\rho(x) - \lambda(y + e^\top x) \\
\text{s.t.} \;& x = x^0 + x^b - x^s, \\
& y + e^\top x + c(x) = 1, \\
& x^b \geq 0,\ x^s \geq 0.
\end{aligned}
$$

Without transaction costs, $y + e^\top x = 1$, (P_2) reduces to the traditional mean-risk portfolio optimization problem. While in the presence of transaction costs, the objective function may be explained by minimizing the portfolio risk $\rho(x)$ and maximizing the actual investment $y + e^\top x$ and the expected return $r_f y + \bar{r}^\top x$.

(P_2) has a drawback that in the absence of a risk-free asset and $\lambda = 0$, there may usually exist an untractable optimal portfolio where simultaneously buying and selling a particular asset may take place. To deal with this dilemma, Mitchell and Braun (2004) propose a scaled risk measurement that replaces the objective function in (P_2) by $\frac{\rho(x)}{(y+e^\top x)^2}$. They prove that the scaled risk measure can be formulated as an equivalent convex program by fractional programming techniques if the transaction costs are linear, piecewise linear or quadratic.

In the discussion thus far, there is an implicit assumption that the transaction costs are paid when the investor rebalances his portfolio. An alternative model for portfolio revision in the literature is that the transaction costs are assumed to be paid at the end of the planning period (Yoshimoto 1996; Li et al. 2000; Best and

Hlouskova 2005). In this case, the investor can make the decision according to the following optimization problem:

$$
(P_3) \qquad \begin{aligned} \min_{x \in X} \ & -(r_{\mathrm{f}} y + \bar{r}^\top x - c(x)) + \gamma \rho(x) \\ \text{s.t.} \ & x = x^0 + x^{\mathrm{b}} - x^{\mathrm{s}}, \\ & y + e^\top x = 1, \\ & x^{\mathrm{b}} \geq 0, \ x^{\mathrm{s}} \geq 0. \end{aligned} \tag{6.7}
$$

(P_3) differs from (P_2) in the objective function and constraint (6.7). It is important to note that $c(x)$ in (P_3) can be also regarded as some sort of penalty due to changes in holdings (Sharpe 1987). Because the transaction costs are paid at the end of the planning period, we may regard $r_{\mathrm{f}} y + \bar{r}^\top x - c(x)$ as the expected return that is not involved in the budget constraint. On the other hand, since $y = 1 - e^\top x$, (P_3) can be rewritten as

$$
(P_4) \qquad \begin{aligned} \min_{x \in X} \ & -(\bar{r} - r_{\mathrm{f}} e)^\top x + c(x) + \gamma \rho(x) \\ \text{s.t.} \ & x = x^0 + x^{\mathrm{b}} - x^{\mathrm{s}}, \\ & e^\top x \leq 1, \\ & x^{\mathrm{b}} \geq 0, \ x^{\mathrm{s}} \geq 0. \end{aligned}
$$

Let $(x^*, x^{\mathrm{b}*}, x^{\mathrm{s}*})$ be an optimal solution to (P_4). Then the amount invested in the risk-free asset is $y^* = 1 - e^\top x^*$. Moreover, if the investor can borrow the risk-free asset freely, budget constraint (6.7) can also be removed, and hence the constraint $e^\top x \leq 1$ in (P_4).

In summary, no matter whether the transaction costs are paid at the beginning or end of the planning horizon, the portfolio optimization problems (P_2) and (P_4) will reduce to the traditional mean-risk model if the costs are ignored by the investor.

6.4 Mean–Variance Framework

Following Chen et al. (1971), we summarize the portfolio revision problem in the mean–variance framework as

$$
(P_5) \qquad \begin{aligned} \min_{x \in X} \ & -(r_{\mathrm{f}} y + \bar{r}^\top x) + \gamma x^\top \Sigma x - \lambda (y + e^\top x) \\ \text{s.t.} \ & x = x^0 + x^{\mathrm{b}} - x^{\mathrm{s}}, \\ & y + e^\top x + c(x) \leq 1, \\ & x^{\mathrm{b}} \geq 0, \ x^{\mathrm{s}} \geq 0, \end{aligned}
$$

if the transaction costs are paid at the beginning of the planning horizon, or

$$
(P_6) \qquad \begin{aligned} \min_{x \in X} \ & -(\bar{r} - r_{\mathrm{f}} e)^\top x + c(x) + \gamma x^\top \Sigma x \\ \text{s.t.} \ & x = x^0 + x^{\mathrm{b}} - x^{\mathrm{s}}, \\ & e^\top x \leq 1, \\ & x^{\mathrm{b}} \geq 0, \ x^{\mathrm{s}} \geq 0, \end{aligned}
$$

if the transaction costs are paid at the end of the planning period. If X is convex, then both (P_5) and (P_6) are convex programs, which can be efficiently solved with interior algorithms.

In the following, we investigate two special cases where a closed-form solution can be obtained from (P_6), and explore the impact of transaction costs on the optimal portfolio position. Notice that the solution of (P_5) can be similarly obtained in both cases and we omit it here (See Chen et al. (1971) for instance).

6.4.1 Analytical Results in the Case of One Risky Asset and One Risk-Free Asset

Suppose there are two assets in the portfolio: one risky asset and one risk-free asset with initial amount denoted by x^0 and y^0 ($x^0 + y^0 = 1$), respectively. The risky asset has mean $\bar{r}$ and variance σ^2, whereas the risk-free asset has constant return rate r_{f} and variance zero. As analyzed above, it is never optimal to have $x^{\mathrm{b}} > 0$ and $x^{\mathrm{s}} > 0$ at the same time, because that would incur both c^{b} and c^{s} on the round-trip. We assume the optimal strategy calls for buying some risky asset, i.e., $x^{\mathrm{b}} > 0$ and $x^{\mathrm{s}} = 0$. In this case, an risk-averse investor's strategy from (P_6) can be integrated into the following optimization problem

$$
\begin{aligned} \min \ & -(\bar{r} - r_{\mathrm{f}})x + c^{\mathrm{b}} x^{\mathrm{b}} + \gamma \sigma^2 x^2 \\ \text{s.t.} \ & x = x^0 + x^{\mathrm{b}}. \end{aligned}
$$

To find the optimal solution, we remove x and write its first-order necessary condition with respect to x^{b} as

$$
-(\bar{r} - r_{\mathrm{f}}) + c^{\mathrm{b}} + 2\gamma\sigma^2 (x^0 + x^{\mathrm{b}}) = 0.
$$

Then, we have the optimal solution

$$
x^* = x^0 + x^{\mathrm{b}*} = \frac{\bar{r} - r_{\mathrm{f}}}{2\gamma\sigma^2} - \frac{c^{\mathrm{b}}}{2\gamma\sigma^2}.
$$

Obviously, in the buying case, there are two parts: the first one is the optimal position (with the highest Sharpe ratio), $\frac{(\bar{r}-r_{\mathrm{f}})}{2\lambda\sigma^2}$, without considering the transaction costs, and the second term is the amount resulting from the transaction costs for buying the risky asset.

Now we consider the contrary case where the optimal strategy is to sell some risky asset, i.e., $x^{\mathrm{b}} = 0$ and $x^{\mathrm{s}} > 0$. Then the portfolio optimization problem (P_6) reduces to

$$\begin{aligned} \min \ & -(\bar{r} - r_{\mathrm{f}})x + c^{\mathrm{s}}x^{\mathrm{s}} + \gamma\sigma^2 x^2 \\ \text{s.t.} \ & x = x^0 - x^{\mathrm{s}}. \end{aligned}$$

According to optimization theory, we have the optimal position invested in the risky asset as

$$x^* = x^0 - x^{\mathrm{s}*} = \frac{\bar{r} - r_{\mathrm{f}}}{2\gamma\sigma^2} - \frac{c^{\mathrm{s}}}{2\gamma\sigma^2}.$$

Similar to the buying case, the optimal solution also has two terms: the first one is the optimal solution without considering transaction costs, and the second term is the impact of transaction costs incurred from portfolio revision. Generally, when considering the transaction costs, the amount invested in the risky asset is always no more than that in the case of no transaction costs.

It should be mentioned that we may generalize and expand the risky asset as the market portfolio (or the market index), which has the highest Sharpe ratio. In this case, all investors with different risk aversion λ hold a mix of the market portfolio and the risk-free asset according to the two-fund separation theorem. The portfolio revision may be the result of a change in the investor's risk aversion. For instance, when the investor becomes more progressive/optimistic, he will increase the position invested in the market portfolio. This corresponds to the strategy (P_2), where c^{b} is the percent transfer cost of buying the market portfolio. Conversely, when the investor becomes more conservative/pessimistic, he will decrease the position in the market portfolio, resulting in selling transfer cost c^{s}, corresponding to (P_3).

6.4.2 Analytical Results in the Case of Two Risky Assets

Now we consider the case where two risky assets are assumed, respectively, to have random returns r_1 and r_2, with means of $\bar{r}_1$ and $\bar{r}_2$, and variance of σ_1^2 and σ_2^2. The correlation between these two assets is assumed to be ρ. The composition of the optimal portfolio is the solution of the maximization problem below, assuming symmetric transaction costs, $c_j^{\mathrm{b}}(x) = c_j^{\mathrm{s}}(x) = c_j|x|$; and further, the optimal strategy calls for buying some asset 1 and selling some asset 2, i.e., $x_1^{\mathrm{b}} = x_2^{\mathrm{s}} > 0$:

$$\begin{aligned}\min_{x_1,x_2} \quad & -(\bar{r}_1 x_1 + \bar{r}_2 x_2) + (c_1 x_1^{\mathrm{b}} + c_2 x_2^{\mathrm{s}}) + \gamma(x_1^2\sigma_1^2 + x_2^2\sigma_2^2 + 2\rho\sigma_1\sigma_2 x_1 x_2) \\ \text{s.t.} \quad & x_1^{\mathrm{b}} = x_1 - x_1^0,\ x_2^{\mathrm{s}} = x_2^0 - x_2, \\ & x_1 + x_2 = 1.\end{aligned} \tag{6.8}$$

It should be noted that for the special case of no transaction costs, i.e., $c_1 = c_2 = 0$, problem (6.8) reduces to the classical mean–variance optimization model. Solving (6.8) yields the following solution:

$$x_1^* = \frac{\gamma\sigma_2(\bar{r}_1 - c_1) + \rho\sigma_1(\bar{r}_2 + c_2)}{2\sigma_1^2\sigma_2(\gamma^2 + \rho^2)},$$

$$x_2^* = \frac{\rho\sigma_2(\bar{r}_1 - c_1) + \gamma\sigma_1(\bar{r}_2 + c_2)}{2\sigma_1\sigma_2^2(\gamma^2 + \rho^2)}.$$

Strictly speaking, we should also develop the reverse case, namely, to sell asset 1 and buy asset 2. The solution is entirely symmetric to (x_1^*, x_2^*) and is found simply by interchanging subscripts there. In addition, there is also the case of an investor not changing his portfolio at all, that is, $x_1 = x_1^0$ and $x_2 = x_2^0$.

6.5 Extension

In practice, the mean return and the covariance matrix are typically estimated based on empirical return samples. These estimated values typically have large errors, particularly for the mean returns. The mean–variance model can be very sensitive to the estimation error in mean return: small differences in the estimate of $\bar{r}$ can result in large variations in the optimal portfolio composition. To protect the performance against estimation risk and to alleviate the sensitivity of the mean–variance model to uncertain input estimates, min–max robust optimization yields the portfolio, which has the best performance under the worst-case scenarios. As demonstrated by Broadie (1993), because the covariance matrix can typically be estimated more accurately than the mean return, in this chapter, we ignore robust consideration with respect to Σ.

We first consider the min–max robust mean–variance portfolio:

$$\begin{aligned}\min_{x\in X} \ \max_{r} \ & -r^\top x + \gamma x^\top \Sigma x \\ \text{s.t.} \ & (r - \bar{r})^\top \Sigma^{-1}(r - \bar{r}) \le \chi^2,\end{aligned} \tag{6.9}$$

where γ is the risk-aversion parameter and χ^2 is a nonnegative number.

For any feasible $x \in X$, let us look at the following optimization problem with respect to r:

$$\begin{aligned} \min_{r \in \Re^n} \quad & r^\top x \\ \text{s.t.} \quad & (r - \bar{r})\Sigma^{-1}(r - \bar{r}) \leq \chi^2. \end{aligned}$$

Apparently, the objective is linear, the constraint is quadratic, and hence we can use the necessary and sufficient Kuhn-Tucker condition to obtain its optimal solution

$$r^* = \bar{r} - \frac{\chi}{\sqrt{x^\top \Sigma x}}$$

with unique optimal objective value

$$\bar{r}^\top x - \chi\sqrt{x^\top \Sigma x}.$$

Therefore, problem (6.9) reduces to the following second-order cone program:

$$\min_{x \in X} -\bar{r}^\top x + \chi\sqrt{x^\top \Sigma x} + \gamma x^\top \Sigma x.$$

Since this is a convex programming problem, it is easy to show that there exist $\tilde{\chi} \geq 0$ such that the above problem is equivalent to

$$\begin{aligned} \min_{x \in X} \quad & -\bar{r}^\top x + \gamma x^\top \Sigma x \\ \text{s.t.} \quad & \sqrt{x^\top \Sigma x} \leq \tilde{\chi}, \end{aligned}$$

which is further equivalent to

$$\begin{aligned} \min_{x \in X} \quad & -\bar{r}^\top x + \gamma x^\top \Sigma x \\ \text{s.t.} \quad & x^\top \Sigma x \leq \tilde{\chi}^2. \end{aligned}$$

According to the convexity of the problem and the Kuhn-Tucker conditions, there exists a $\hat{\gamma} \geq 0$ such that the above problem is equivalent to

$$\min_{x \in X} \quad -\bar{r}^\top x + \tilde{\gamma} x^\top \Sigma x,$$

where $\tilde{\gamma} = \gamma + \hat{\gamma}$ (interested reader can refer to Zhu et al. (2007) for a detailed proof). Up to now, we obtain a surprising conclusion that the robust mean–variance approach cannot improve the portfolio performance, and instead, it has the same efficient frontiers as the classical mean–variance model. Mulvey et al. (2008) use a momentum strategy to empirically compare several extended mean–variance models and conclude that "the traditional Markowitz model has performed reasonably well when compared with its robust versions."

6.6 Practical Application

In this section, we present an illustrative example to demonstrate the impact of transaction costs on the optimal portfolio to be rebalanced. Particularly, we consider the problem of portfolio revision using 10 industry portfolios with equal initial endowment (as mentioned in Sect. 6.2, we normalize the initial portfolio and hence $x_j^0 = 0.1$ for $j = 1, \cdots, 10$).[1] The components of the 10 industry portfolio are given in Table 6.1. Daily average value weighted returns are considered from January 1, 2000 to December 31, 2007, including 2010 observations. Table 6.2 exhibits the expected return and covariances for daily, monthly, and year returns (the data for yearly return ranges from 1978 to 2007).

We consider the impact of transaction costs on the efficient frontier of mean–variance optimal portfolios. More specifically, we assume the transaction costs are paid at the end of the planning horizon, i.e., the optimal portfolios can be obtained by solving problems (P_6). In this example, the investor is assumed to revise his portfolio daily; in the morning, he revises his portfolio according to problem (P_6), and in the evening, he pays the transaction costs. The feasible portfolio set is assumed as $X = \{x_i \geq -0.2 : x \in \Re^{10}\}$, i.e., short position is allowed but has a lower bound -0.2. To investigate the impact of transaction costs, we explore the property of efficient frontiers from which a rational investor will choose his optimal portfolio. From Fig. 6.1, transaction costs are obviously necessary to be taken into account when employing an active portfolio trading strategy. With two cases, $c_j^{\mathrm{b}} = c_j^{\mathrm{s}} = 0$

Table 6.1 Details for 10 industry portfolios

1	NoDur	Consumer NonDurables: Food, Tobacco, Textiles, Apparel, Leather, Toys
2	Durbl	Consumer Durables: Cars, TV's, Furniture, Household Appliances
3	Manuf	Manufacturing: Machinery, Trucks, Planes, Chemicals, Off Furn, Paper, Com Printing
4	Enrgy	Oil, Gas, and Coal Extraction and Products
5	HiTec	Business Equipment: Computers, Software, and Electronic Equipment
6	Telcm	Telephone and Television Transmission
7	Shops	Wholesale, Retail, and Some Services (Laundries, Repair Shops)
8	Hlth	Healthcare, Medical Equipment, and Drugs
9	Utils	Utilities
10	Other	Other: Mines, Constr, BldMt, Trans, Hotels, Bus Serv, Entertainment, Finance

[1] The data are available from the website of Professor Kenneth French http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data-library.htm.

Table 6.2 Summary statistics of portfolio

NoDur	Durbl	Manuf	Enrgy	HiTec	Telcm	Shops	Hlth	Utils	Other
Daily expected return (%)									
0.0703	0.0406	0.0834	0.1323	0.0654	–0.0009	0.0752	0.0987	0.0616	0.0718
Daily covariance (%)									
0.5247	0.6235	0.6078	0.4849	0.8375	0.8350	0.6035	0.6472	0.4021	0.4706
0.6235	1.0454	0.8534	0.6611	1.2276	1.2003	0.8378	0.9076	0.5156	0.6503
0.6078	0.8534	0.9157	0.7856	1.2226	1.1787	0.8141	0.8884	0.5458	0.6444
0.4849	0.6611	0.7856	1.9004	0.8622	0.8339	0.6211	0.7511	0.6840	0.5219
0.8375	1.2276	1.2226	0.8622	2.5224	2.2632	1.2219	1.6304	0.5797	0.9279
0.8350	1.2003	1.1787	0.8339	2.2632	2.5958	1.1892	1.5169	0.6122	0.9183
0.6035	0.8378	0.8141	0.6211	1.2219	1.1892	0.8809	0.8909	0.4823	0.6342
0.6472	0.9076	0.8884	0.7511	1.6304	1.5169	0.8909	1.4530	0.5137	0.6951
0.4021	0.5156	0.5458	0.6840	0.5797	0.6122	0.4823	0.5137	0.8215	0.4212
0.4706	0.6503	0.6444	0.5219	0.9279	0.9183	0.6342	0.6951	0.4212	0.5285
Monthly expected return (%)									
0.8575	0.3498	0.8583	1.5302	–0.1806	–0.4526	0.3070	0.3702	1.1595	0.5286
Monthly covariance (%)									
10.1767	6.6563	7.6224	6.5753	4.9721	6.9501	7.6918	5.4699	7.0175	7.9088
6.6563	39.3862	19.7554	11.7425	30.2980	20.6187	18.6279	5.2806	8.5486	17.7803
7.6224	19.7554	17.6824	12.6521	22.8372	13.5418	13.7950	5.8287	7.3997	13.7399
6.5753	11.7425	12.6521	29.6173	12.0461	9.1655	7.4645	4.6172	14.7046	10.3207
4.9721	30.2980	22.8372	12.0461	76.4744	34.8299	21.6430	10.5466	2.9134	20.8838
6.9501	20.6187	13.5418	9.1655	34.8299	34.6161	15.8662	8.3762	3.6022	15.1647
7.6918	18.6279	13.7950	7.4645	21.6430	15.8662	19.9695	3.9989	5.1589	14.2172
5.4699	5.2806	5.8287	4.6172	10.5466	8.3762	3.9989	13.7028	6.7478	6.2923
7.0175	8.5486	7.3997	14.7046	2.9134	3.6022	5.1589	6.7478	20.7251	9.5173
7.9088	17.7803	13.7399	10.3207	20.8838	15.1647	14.2172	6.2923	9.5173	17.3139
Yearly expected return									
0.1437	0.1272	0.1650	0.2075	0.2178	0.2086	0.1521	0.2302	0.1603	0.1804
Yearly covariance									
0.0313	0.0335	0.0264	0.0058	0.0328	0.0255	0.0384	0.0446	0.0120	0.0330
0.0335	0.0472	0.0326	0.0092	0.0528	0.0460	0.0470	0.0514	0.0094	0.0410
0.0264	0.0326	0.0339	0.0386	0.0504	0.0395	0.0360	0.0495	0.0108	0.0354
0.0058	0.0092	0.0386	0.1944	0.0574	0.0261	0.0116	0.0481	0.0142	0.0318
0.0328	0.0528	0.0504	0.0574	0.1593	0.1266	0.0585	0.1136	0.0047	0.0558
0.0255	0.0460	0.0395	0.0261	0.1266	0.1469	0.0428	0.0666	0.0064	0.0415
0.0384	0.0470	0.0360	0.0116	0.0585	0.0428	0.0560	0.0637	0.0105	0.0446
0.0446	0.0514	0.0495	0.0481	0.1136	0.0666	0.0637	0.1563	0.0218	0.0569
0.0120	0.0094	0.0108	0.0142	0.0047	0.0064	0.0105	0.0218	0.0201	0.0128
0.0330	0.0410	0.0354	0.0318	0.0558	0.0415	0.0446	0.0569	0.0128	0.0474

The return series are available from Professor Kenneth French's website

and $c_j^{\mathrm{b}} = c_j^{\mathrm{s}} = 2.5\%$, the efficient frontier are dramatically lowered by the transaction costs in a nonlinear pattern. We do not include the risk-free asset into the portfolio since moving the cash in and out of the account does not incur any cost in practice.

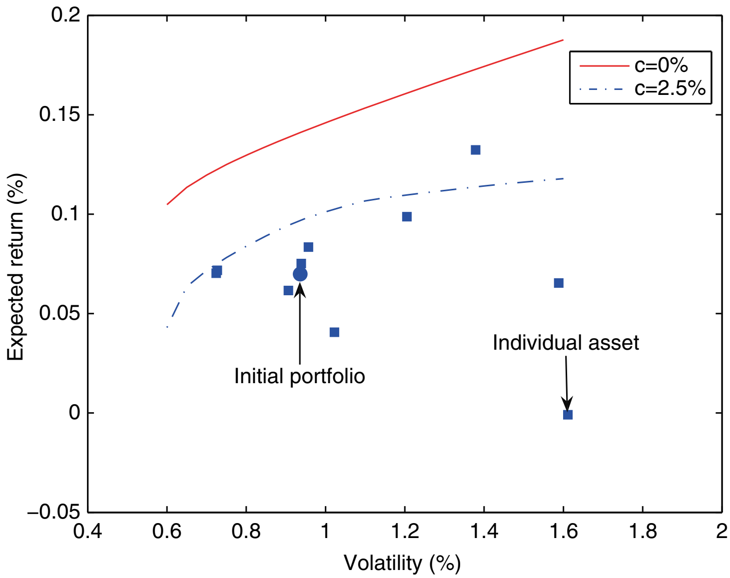


Fig. 6.1 Efficient frontier of mean–variance strategy with daily revision in the presence of symmetric transaction costs $c_j^{\mathrm{b}} = c_j^{\mathrm{s}} = 0\%$ and 2.5%. The investor solves problem (P_6) by assuming that the transaction costs are paid at the end of the horizon. The graphs use data of 10 industry portfolios from January 1, 2000 to December 31, 2007

It is important to note that if the risk aversion coefficient γ is very large, the risk term in (P_6) dominates the expected return term. In the extreme case, (P_6) reduces to the risk minimization problems that do not pay attention to the expected return $\bar{r}^{\top}x - c(x)$, and hence the investor may rebalance some assets that have lower daily returns than 2.5% or 5%, yielding a negative return. The smaller the risk, the larger the impact of transaction costs. That is, the effect of transaction costs becomes more significant as the risk the investor would like to take on decreases.

On the other hand, the blue-circle dot in Fig. 6.1 represents the initial portfolio. When the transaction costs increase until the efficient frontier touches the blue-circle dot, there will be no revision if the investor has a risk tolerance equal to the initial portfolio. After that, the initial portfolio will be optimal, and there will be no revision in the sense of Dybvig (2005).

Although the impact magnitude of the transaction costs may be dataset-specific, its effect for the 10 industry portfolios is very huge, or even unacceptable if there is no risk-free asset. From Fig. 6.1, the discrepancy of expected returns between optimal portfolio without or with 2.5% transaction costs is close to 0.05%. That means, the investor should pay 0.05% of his investment for daily rebalance, resulting in a 12.5% yearly cost (250 trading days). The discrepancy of the efficient frontier becomes larger as the volatility increases, i.e., the investor becomes less risk averse (γ decreases) and hence accepts a larger risk for higher expected return.

In a graphical representation of volatility vs. expected return, optimal portfolios without out transaction costs should be found to the left or above to the individual volatility and expected return due to diversification. However, this is not true when the transaction costs are incorporated into the portfolio optimization problem. As the transaction costs increase from 0 to 2.5%, there is one asset lying above the efficient frontier (Fig. 6.1).

In practice, because the daily revision is costly, investors usually reduce transaction costs by decreasing the revision frequency. To investigate the difference of transaction costs' effect in the revision frequencies, Figures 6.2 and 6.3 exhibit the efficient frontiers for the 10 industry portfolios with monthly and yearly rebalance, respectively. The monthly data ranges from January 2000 to December 2007 resulting in 96 samples, and the yearly data ranges from 1978 to 2007, resulting in 30 samples. Similar to the daily rebalance, the lower bound of optimal portfolio is −0.2, i.e., any short position is no more than 20% of the total portfolio market value. In this example, we do not consider the estimation error and hence regard the estimation from the sample as the true parameters of the portfolio. Although the impact of transaction costs decreases as the revision frequency decreases (the cumulative impact with daily rebalance is larger than that of monthly rebalance and the monthly is larger than the yearly), it is still significant. Again, there are 3 assets out of 10 whose coordinates lie above the efficient frontier when the transaction

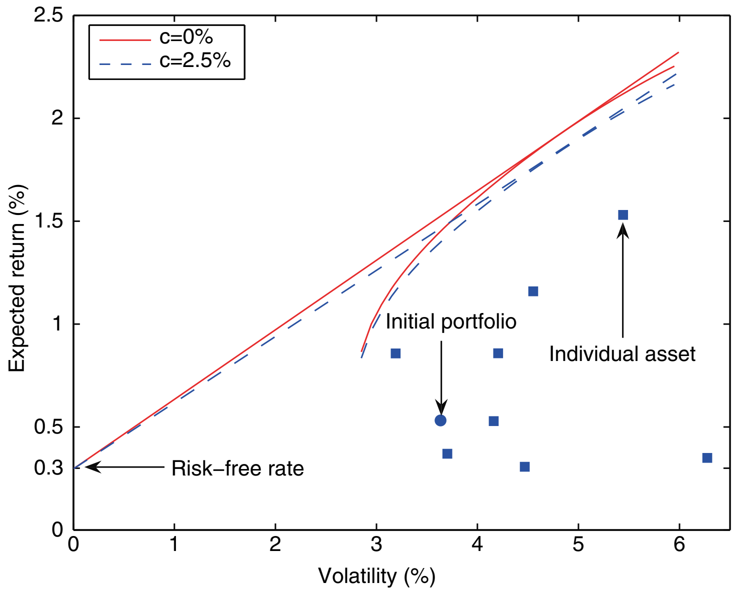


Fig. 6.2 Efficient frontier of mean–variance strategy with monthly revision in the presence of symmetric transaction costs $c_j^{\mathrm{b}} = c_j^{\mathrm{s}} = 0\%$ and 2.5%. The investor solves problem (P_6) by assuming that the transaction costs are paid at the end of the horizon. We assume that the risk-free rate of return is 3.6% annually and hence 0.3% monthly. Each square corresponds to one individual standard deviation (volatility) and its expected return. The graphs use data of 10 industry portfolios from January 2000 to December 2007. The *squares* of the 5th and 6th assets lie outside the [0, 6.5; 0, 2.5] range for negative means

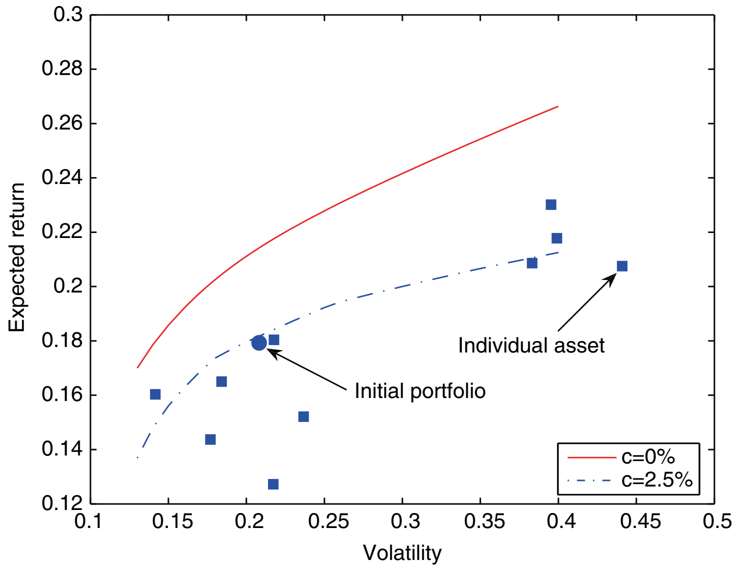


Fig. 6.3 Efficient frontier of mean–variance strategy with yearly revision in the presence of symmetric transaction costs $c_j^b = c_j^s = 0\%$ and 2.5%. The investor solves problem (P_6) by assuming that the transaction costs are paid at the end of the horizon. Each square corresponds to one individual standard deviation (volatility) and its expected return. The graphs use data of 10 industry portfolios from 1978 to 2007

costs are 2.5%. In summary, when seeking to develop new methodology to improve portfolio performance, transaction costs involved in the revision process should be an important factor to be considered, especially for those active portfolio managers.

6.7 Conclusion

In the revision process, there is always a transaction cost associated with buying and selling an asset due to brokerage fees, bid-ask spreads, or taxes. In this chapter, we consider the problem of portfolio revision with transaction costs, which are paid either at the beginning of the planning horizon or at the end of the planning horizon. We demonstrate that the impact of transaction costs can be integrated into the classical mean–variance framework and show that even some analytical solutions under mild assumptions can be obtained via optimization techniques.

While the risk measure used in this chapter is the variance, there are alternative risk measures that have been suggested by theoreticians and market practitioners. The most popular measure is conditional value at risk (CVaR). Chen et al. (2008) show how the mean–variance portfolio revision model presented in this chapter can be applied when CVaR is the risk measure.

References

Adcock, C.J. and N. Meade (1994) A simple algorithm to incorporate transaction costs in quadratic optimization, *European Journal of Operational Research*, **79**: 85–94.

Best, M.J. and R.R. Grauer (1991) On the sensitivity of mean–variance efficient portfolios to changes in asset means: some analytical and computional results, *Review of Financial Studies*, **4**(2): 315–342.

Best, M.J. and J. Hlouskova (2005) An algorithm for portfolio optimization with transaction costs, *Management Science*, **51**(11): 1676–1688.

Black, F. and R. Litterman (1992) Global portfolio optimization, *Financial Analysts Journal*, **48**: 28–43.

Bloch, M., J.B. Guerard, H.M. Markowitz, P. Todd and G.L. Xu (1993) A comparison of some aspects of the U.S. and Japanese equity markets, *Japan & the World Economy*, **5**: 3–26.

Bloomfield, T., R. Leftwich and J.B. Long (1977) Portfolio strategies and performance, *Journal of Financial Economics*, **12**(2): 201–218.

Brandt, M.W. (1999) Estimating portfolio and consumption choice: a conditional Euler equation approach, *Journal of Finance*, **54**: 1609–1645.

Broadie, M. (1993) Computing efficient frontiers using estimated parameters, *Annals of Operations Research*, **45**: 21–58.

Campbell, J.Y. and L.M. Viceira (1999) Consumption and portfolio decisions when expected returns are time varying, *Quarterly Journal of Economics*, **114**: 433–495.

Chan, Louis K.C., N. Jegadeesh and J. Lakonishok (1996) Momentum strategies, *Journal of Finance*, **51**: 1681–1713.

Chan, Louis K.C., N. Jegadeesh, and J. Lakonishok (1999) The profitability of momentum strategies, *Financial Analysts Journal*, **99**: 80–90.

Chen, A.H., F.J. Fabozzi and D. Huang (2008) Portfolio revision under mean-variance and mean-CVaR with transaction costs, Working paper.

Chen, A.H., F.C. Jen and S. Zionts (1971) The optimal portfolio revision policy, *Journal of Business*, **44**: 51–61.

Chopra, V.K. (1993) Mean-variance revisited: near-optimal portfolios and sensitivity to input variations, *Journal of Investing*, **2**(1): 51–59.

Chopra, V.K. and W.T. Ziemba (1993) The effects of errors in means, variances, and covariances on optimal portfolio choices, *Journal of Portfolio Management*, **19**(2): 6–11.

Davis, M.H.A. and A.R. Norman (1990) Portfolio selection with transaction costs, *Mathematics of Operations Research*, **15**(4): 676–713.

Dybvig, P.H. (2005) Mean-variance portfolio rebalancing with transaction costs, Working paper, Washington University in Saint Louis.

Fabozzi, F.J., F. Gupta and H.M. Markowitz (2002) The legacy of modern portfolio, *Journal of Investing*, **11**(3): 7–22.

Fama, E.F. and K.R. French (1992) Cross-sectional variation in expected stock returns, *Journal of Finance*, **47**: 427–465.

Fama, E.F. and K.R. French (1995) Size and the book-to-market factors in earnings and returns, *Journal of Finance*, **50**: 131–155.

Grinold, R. and R. Kahn (1999) *Active portfolio management*, New York: McGraw-Hill/Irwin.

Jacobs, B.I., K.N. Levy and H.M. Markowitz (2005) Portfolio optimization with factors, scenarios, and realistic short positions, *Operations Research*, **53**: 586–599.

Johnson K.H. and D.S. Shannon (1974) A note on diversification and the reduction of dispersion, *Journal of Financial Economics*, **1**(4): 365–372.

Konno H. and A. Wijayanayake (2001) Portfolio optimization problem under concave transaction costs and minimal transaction unit constraints, *Mathematical Programming*, **89**(2): 233–250.

Krokhmal, P., S. Uryasev and J. Palmquist (2001) Portfolio optimization with conditional value-at-risk objective and constraints, *Journal of Risk*, **4**(2): 43–68.

Li, Z.F., S.Y. Wang and X.T. Deng (2000) A linear programming algorithm for optimal portfolio selection with transaction costs, *International Journal of Systems Science*, **31**(1): 107–117.

Lobo, M.S., M. Fazel and S. Boyd (2007) Portfolio optimization with linear and fixed transaction costs, *Annals of Operatons Research*, **152**(1): 341–365.

Markowitz, H.M. (1976) Investment in the long run: new evidence for an old rule, *Journal of Finance*, **31**: 1273–1286.

Markowitz, H.M. (1999) The early history of portfolio theory: 1600-1960, *Financial Analysts Journal*, **99**: 5–16.

Markowitz, H.M. (2000) *Mean-variance analysis in portfolio choice and capital markets*, New Hope, PA: Frank J. Fabozzi Associates.

Michaud, R. (1998) *Efficient asset management*, New York: Oxford University Press. First published by Harvard Business School Press.

Mitchell, J.E. and S. Braun (2004) Rebalancing an investment portfolio in the presence of convex transaction costs, Working paper.

Mossin, J. (1968) Optimal multiperiod portfolio policies, *Journal of Business*, **41**: 215–229.

Mulvey, J.M., W.C. Kim and M. Bilgili (2008) Liking momentum strategies with single-period portfolio models, to appear in the *Handbook of portfolio construction: contemporary applications of Markowitz techniques*, edited by J. Guerard, New York: Springer.

Rachev, S.T., C. Menn and F.J. Fabozzi (2005) *Fat-tailed and skewed asset return distributions: implications for risk management, portfolio selection and option pricing*, Wiley.

Rosenberg, B. (1974) Extra-market components of covariance in security returns, *Journal of Financial and Quantitative Analysis*, **9**: 263–274.

Rosenberg, B. and V. Marathe (1979) Tests of capital asset pricing hypotheses, in *Research in Finance 1*, edited by H. Levy, Greenwich, Conn.: JAI Press.

Rubinstein M. (2002) Markowitz's "Portfolio Selection": a fifty-year retrospective, *Journal of Finance*, **57**(3): 1041–1045.

Rudd, A. and B. Rosenberg (1979) Realistic portfolio optimization, in *Portfolio theory, 25 years after*, edited by E. Elton and M. Gruber, Amsterdam: North-Holland Publishing Co.

Rudd, A. and B. Rosenberg (1980) The 'market model' in investment management, *Journal of Finance*, **35**: 597–607.

Sharpe, W.F. (1971) A linear programming approximation for the general portfolio problem, *Journal of Financial and Quantitative Analysis*, **6**: 1263–1275.

Sharpe, W.F. (1984) Factor models, CAPMs, and the APT, *Journal of Portfolio Management*, **11**: 21–25.

Sharpe, W.F. (1987) An algorithm for portfolio improvement, in *Advances in mathematical programming and financial planning*, edited by K.D. Lawrence, J.B. Guerard, Jr., and Gary D. Reeves, Greenwich: JAI Press, pp: 155–170.

Smith, K.V. (1967) A transition model for portfolio revision, *Journal of Finance*, **22**: 425–439.

Yoshimoto A. (1996) The mean-variance approach to portfolio optimization subject to transaction costs, *Journal of Operations Research Society of Japan*, **39**(1): 99–117.

Zhu, L., T.F. Colema and Y. Li (2007) Min-max robust and CVaR robust mean-variance portfolios, Working paper, David R. Cheriton School of Computer Science, University of Waterloo.

Chapter 7
Principles for Lifetime Portfolio Selection: Lessons from Portfolio Theory

James H. Vander Weide

7.1 Introduction

An individual's savings and investment choices at various stages of life are among the most important decisions he or she can make. A person entering the workforce in 2008 can expect to work for approximately 40–45 years and to live in retirement for an additional 20–25 years. During his working life, an individual must accumulate sufficient assets not only to live comfortably in both good and bad economic times, but also to live comfortably in retirement. To achieve the goal of maximizing economic welfare over his expected lifetime, an individual consumer/investor should have a sound understanding of the basic economic principles of lifetime portfolio selection.

The lifetime consumption/investment decision problem is complex. Suppose an individual consumer/investor divides her expected remaining lifetime into N equal periods. At the beginning of period 1, she must allocate her wealth W_0 to consumption C_0 and investment W_0-C_0. Her wealth at the beginning of the next period, W_1, will depend on both the amount she chooses to invest at the beginning of period 1 and the return she earns on her investment. The consumer/investor recognizes that she will continue to make consumption/investment decisions at the beginning of each period of her life. Her goal is to maximize her expected utility from lifetime consumption. Since expected utility from lifetime consumption depends on consumption/investment decisions in every period of her life, and the opportunities in later periods depend on the results of decisions in earlier periods, the individual consumer/investor must potentially solve a complex N period optimization problem simply to make the correct consumption/investment decision at the beginning of period 1.

Portfolio theory is concerned with developing general principles and practical models for making sound lifetime portfolio decisions. Much of the current research on portfolio theory emanates from the path-breaking mean–variance portfolio model of Nobel Laureate Harry Markowitz. Markowitz (1952, 1959) recommends that in

J.H. Vander Weide (✉)
Fuqua School of Business, Duke University, Durham, NC, USA
e-mail: jim.vanderweide@duke.edu

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_7,
© Springer Science+Business Media, LLC 2010

making investment decisions, investors should explicitly recognize investment risk as measured by variance of return, as well as expected return. He describes how the variance of return on a portfolio of securities depends on the amount invested in each security, the variance of return on each security, and the correlation between the returns on each pair of securities. He also suggests that investors limit their choices to an efficient set of portfolios that provide the highest mean return for any level of variance and the lowest variance of return for any level of mean. By providing an intuitively appealing measure of portfolio risk and a framework for analyzing the basic risk/return tradeoff of portfolio decisions, Markowitz revolutionized both the theory and practice of portfolio management. For that reason, Markowitz is properly called the father of Modern Portfolio Theory.[1]

The beauty of the Markowitz mean–variance model lies in its blend of elegance and simplicity. Markowitz achieves elegance by providing investors a sophisticated tool for: (1) understanding how portfolio mix decisions affect portfolio risk; and (2) determining those portfolios that provide an efficient combination of risk and return. He achieves simplicity by focusing solely on the economic trade-off between portfolio risk and return in a single-period world.[2]

Although the mean–variance model continues to be the most widely used portfolio model in financial practice, economists have devoted considerable effort to research on two additional models of portfolio behavior, the geometric mean model and the lifetime consumption–investment model. These models offer significant additional insights into optimal portfolio behavior. The purpose of this chapter is to review the major findings of the research literature on the mean–variance model, the geometric mean model, and the lifetime consumption–investment model, and, on the basis of this review, to develop a set of practical guidelines for making lifetime portfolio decisions.

7.2 The Markowitz Mean–Variance Model

Investors make portfolio decisions by selecting the securities to include in the portfolio and the amount to invest in each security. In making risky portfolio choices, the Markowitz mean–variance approach assumes that investors: (1) consider only the mean and variance of the probability distribution of portfolio and security returns; (2) for a given level of mean return, prefer a portfolio with a lower variance of return; and (3) for a given level of variance of return, prefer a portfolio with a higher mean return.

As Markowitz demonstrates, the above assumptions suggest that an investor's portfolio decision problem can be solved in three steps. First, an investor can

[1] This chapter is dedicated to Dr. Markowitz in celebration of his 80th birthday.

[2] Markowitz discusses many of the dynamic economic forces that affect lifetime consumption and investment decisions in the later chapters of Markowitz (1959). However, the economic forces described in this discussion are not incorporated directly in his single-period mean–variance model.

estimate the mean and variance of return on each security and the correlation of returns on each pair of securities. Second, an investor can calculate the mean and variance of return on each feasible portfolio and determine an "efficient frontier" of portfolios that offer the lowest variance of return for any level of mean return and the highest mean return for any level of variance of return. Third, an investor can choose a portfolio on the efficient frontier. This chapter will focus primarily on steps two and three.

7.2.1 Estimating the Mean and Variance of Portfolio Returns

Assume that there are N securities and that an investor allocates the proportion X_i of his wealth to security i. Let R_i denote the return on security i and R_p the return on the portfolio of N securities. Then:

$$R_\mathrm{p} = R_1X_1 + R_2X_2 + \cdots + R_nX_n, \tag{7.1}$$

where R_p and $\mathrm{R}_1 \ldots \mathrm{R}_n$ are random variables.

According to (7.1), the portfolio return, R_p, is a weighted average of the returns on the securities in the portfolio. Formulas for calculating the mean and variance of a weighted sum of random variables are presented in most introductory probability texts. Using these formulas, the mean of the portfolio return, E_p, is given by:

$$E_\mathrm{p} = E_1X_1 + E_2X_2 + \cdots + E_nX_n, \tag{7.2}$$

where $E_1, \ldots, E_n$, are the mean, or expected, returns on the individual securities; and the variance of the portfolio return is given by:

$$V_\mathrm{p} = \sum_i V_i X_i^2 + \sum_i \sum_{j>i} 2C_{ij} X_i X_j, \tag{7.3}$$

where V_i is the variance of return on security i, and C_{ij} is the covariance of returns on security i and security j.

In the Markowitz mean–variance model, investment risk is measured by either the variance of the portfolio return or its equivalent, the standard deviation of portfolio return.[3] The formula for portfolio variance (7.3) can be used to provide insight on how investors can reduce the risk of their portfolio investment. Recall that the covariance of returns on security i and security j can be written as the product of the standard deviation of return on security i, SD_i, the standard deviation

[3] Variance and standard deviation of return are considered to be equivalent measures of risk because the standard deviation is the positive square root of the variance, and the positive square root is an order-preserving transformation. Thus, portfolios that minimize the variance of return for any level of mean return will also minimize the standard deviation of return for any level of mean return.

of the return on security j, SD_j, and the correlation of returns on securities i and j, ρ_{ij}:

$$C_{ij} = \mathrm{SD}_i \times \mathrm{SD}_j \times \rho_{ij}. \tag{7.4}$$

To simplify the analysis, assume that: (1) the variances on all securities are equal to the average security variance, $\overline{V}$; (2) the correlation of returns on all securities i and j are equal to the average correlation of return, $\overline{\rho}$, on securities; and (3) the investor allocates $1/N$ of his wealth to all securities.

Under these assumptions, the variance of return on the portfolio, V_p, can be written as:

$$V_\mathrm{p} = \overline{V}\left(\tfrac{1}{N}\right) + \tfrac{N(N-1)}{N^2}\overline{V}\overline{\rho}. \tag{7.5}$$

The effect of variations in $\overline{V}$, $\overline{\rho}$, and N on portfolio variance, V_p, can be determined by calculating the partial derivative of V_p, with respect to each of these variables:

$$\begin{aligned}
\tfrac{\partial V_\mathrm{p}}{\partial \overline{V}} &= \tfrac{1}{N} + \tfrac{N(N-1)}{N^2}\overline{\rho} > 0 \qquad \text{if} \quad \overline{\rho} \geqslant 0,\\
\tfrac{\partial V_\mathrm{p}}{\partial \overline{\rho}} &= \tfrac{N(N-1)}{N^2}\overline{V} > 0,\\
\tfrac{\partial V_\mathrm{p}}{\partial N} &= -\overline{V}(\tfrac{1}{N^2}) + \overline{V}\overline{\rho}(\tfrac{1}{N^2}) = \overline{V}(\overline{\rho} - 1)(\tfrac{1}{N^2}) \leqslant 0.
\end{aligned} \tag{7.6}$$

These equations indicate that the portfolio variance of return can be reduced in three ways: (1) increasing the number of securities in the portfolio; (2) choosing securities having returns that are less correlated with returns on other securities; and (3) if $\overline{\rho}$ is greater than or equal to zero, choosing securities with low variance or standard deviation of returns.

The formulas for portfolio mean and variance, given by (7.2) and (7.3), require estimates of the mean, E_i, and variance, V_i, of return on each security, as well as the covariance, C_{ij}, of returns on each pair of securities. If there are N securities under consideration, (7.2) and (7.3) require N mean estimates, N variance estimates, and $N\,(N-1)/2$ distinct covariance estimates, for a total of $2N + N(N-1)/2$ estimates. To illustrate, assume that an analyst is considering 200 securities for possible inclusion in a portfolio. Then the analyst must estimate 200 mean values, 200 variance values, and 19,900 covariance values to implement the Markowitz mean–variance model. Without simplification, it is unlikely that the analyst could estimate these inputs cost effectively.

One way to reduce the large number of estimates required to implement the Markowitz mean–variance model is to apply the model to asset classes rather than to individual securities. For example, if the universe of securities is divided into large US stocks, small US stocks, global stocks, emerging market stocks, corporate bonds, long-term US government bonds and Treasury bills, the number of input estimates would be reduced from 20,300 to 35. Given the importance of the asset mix decision and the significant reductions in required estimates obtainable by considering asset categories rather than individual securities, it is not surprising that the Markowitz model is frequently applied to asset categories rather than to individual securities.

Another way to reduce the input requirements of the Markowitz mean–variance model is to make one or more simplifying assumptions about the covariance structure of security returns. For example, if one assumes that: (1) the return on an individual security i is related to the return on a market index via the equation:

$$R_i = \alpha_i + \beta_i R_\mathrm{m} + e_i \quad \text{for} \quad i = 1, 2, \cdots, N, \tag{7.7}$$

where R_i is the return on security i, R_m is the return on the market index, and e_i is a random error term; (2) $E[e_i(R_\mathrm{m} - \overline{R_\mathrm{m}})] = 0$; and (3) $E(e_i \times e_j) = 0$; then the means, variances, and covariances of securities' returns are given by:

$$E_i = \alpha_i + \beta_i E_\mathrm{m}, \tag{7.8}$$

$$V_i = \beta_i^2 V_\mathrm{m} + V_{e_i}, \tag{7.9}$$

$$C_{ij} = \beta_i \beta_j V_\mathrm{m}. \tag{7.10}$$

Substituting (7.8)–(7.10) into (7.2) and (7.3), we obtain the following equations for the mean and variance of return on a portfolio of securities:

$$E_\mathrm{p} = \sum_i X_i \left(\alpha_i + \beta_i E_\mathrm{m}\right), \tag{7.11}$$

$$V_\mathrm{p} = \sum_i {X_i}^2 \beta_i^2 V_\mathrm{m} + \sum_i \sum_{j \neq i} X_i X_j \beta_i \beta_j V_\mathrm{m} + \sum_i {X_i}^2 V_{e_i}. \tag{7.12}$$

To estimate the mean and variance of return on any portfolio then, we need only to estimate the α_i, β_i, and V_{e_i} inputs for each security and the expected return, E_m, and variance of return, V_m, on the market index. Thus, the total number of required estimates has been reduced from $2N + [N \times (N - 1)]/2$ to $3N + 2$. If the analyst is considering 200 securities for possible inclusion in a portfolio, the number of required estimates is reduced from 20,300 to 602.

7.2.2 The Feasible Set of Portfolios and the Efficient Frontier

The feasible set of portfolios is the set of all security allocations $(X_1, \ldots X_N)$ that satisfy the individual's portfolio constraints. An obvious portfolio constraint is that the sum of the proportion of wealth invested in all securities must equal 1. Other typical constraints are that the proportion invested in each security must be nonnegative (i.e., short selling is not allowed) and the investor will not invest more than a certain percentage of wealth in any one security.

The Markowitz mean–variance portfolio model allows an investor to translate all feasible portfolio proportions $(X_1, \ldots X_N)$ into feasible combinations of: (1) expected return and variance of return; or (2) expected return and standard deviation

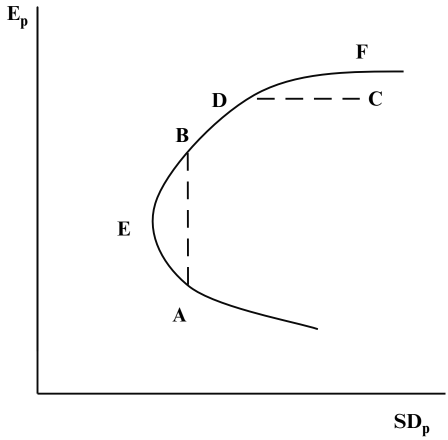


Fig. 7.1 Efficient frontier of risky portfolios

of return. Figure 7.1 shows one such feasible set of E_p, SD_p, combinations. Consider portfolio A shown in Fig. 7.1. Rational Markowitz mean–variance investors would not choose to invest in portfolio A because they could achieve a higher expected return by investing in portfolio B without increasing the portfolio standard deviation. Similarly, rational investors would not invest in portfolio C because they could achieve a lower portfolio standard deviation by investing in portfolio D without sacrificing mean return. The efficient set of portfolios consists of all portfolios with the highest mean return for any given level of standard deviation of return and the lowest standard deviation of return for any given level of mean return. The curved line EBDF is the efficient frontier for the feasible set of risky portfolios shown in Fig. 7.1.

7.2.3 The Effect of a Risk-Free Security on the Shape of the Efficient Frontier

When all securities are risky, the efficient frontier typically has a shape similar to that shown in Fig. 7.1. Suppose now there exists a security with a risk-free rate of return, R_f. Since the risk-free security has zero variance, its mean–standard deviation combination lies on the vertical axis in mean–standard deviation space (see Fig. 7.2). In addition, the risk-free security must lie on the efficient frontier because there are no other securities with the same mean return and a lower standard deviation of return.

To see how the shapes of the feasible set of securities and the efficient frontier change as a result of the presence of a risk-free security, consider a new portfolio consisting of the fraction X invested in a portfolio S of risky securities and a fraction $(1 - X)$ invested in the risk-free security. Because the return on the risk-free

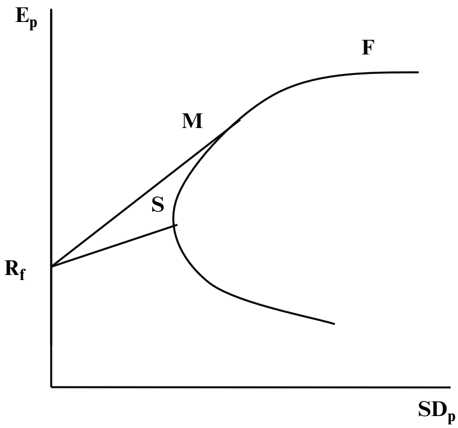


Fig. 7.2 Efficient frontier in presence of risk-free asset

security has zero variance and is uncorrelated with the return on portfolio S, both the mean and standard deviation of return on the new portfolio are linearly related to the mean and standard deviation of return on portfolio S.[4] Thus, the new portfolio lies somewhere on the straight line connecting R_f and S in Fig. 7.2, with its exact location depending on the fraction of wealth X invested in portfolio S.

Since the risky portfolio S in the above example is selected arbitrarily, the revised set of feasible portfolios consists of E_p, SD_p combinations lying on any line connecting R_f with a point such as S in the feasible set of risky portfolios. The slope of such a line is $(E_s - R_f)/SD_s$. Consider now the line connecting R_f with portfolio M in Fig. 7.2. The points on this line represent feasible portfolios with the fraction X invested in the risky portfolio M and the fraction (1-X) invested in the risk-free security. Since the slope $(E_M - R_f)/SD_M$ is greater than the slope $(E_s - R_f)/SD_s$ for any risky portfolio S, for any feasible portfolio not on the line connecting R_f and risky portfolio M, there exists a portfolio on the line connecting R_f and risky portfolio M with a higher mean return and the same standard deviation of return or the same mean return and a lower standard deviation of return. Thus, the points on the line connecting R_f and M are not only feasible but also efficient. Evidently, the new efficient frontier consists of the union of all points on the line connecting R_f and M and all points on the efficient frontier of risky securities between M and F.[5]

---

[4] Specifically, $E_p = X \cdot E_s + (1 - X)R_f$ and $SD_p = X \cdot SD_s$.

[5] This conclusion strictly applies only when the investor cannot finance risky investments with borrowing. If the investor can borrow as well as lend at the risk-free rate, the efficient frontier will consist of the entire straight line emanating from R_f and extending through the tangency point M on the efficient frontier of risky securities.

7.2.4 Identifying the Efficient Frontier

As noted above, the efficient set of portfolios consists of all portfolios with the highest mean return for a given level of variance or standard deviation and the lowest variance (standard deviation) for a given level of mean. Once the E_{i}, V_{i}, and C_{ij} inputs have been estimated, the analyst can calculate the mean–variance efficient frontier by solving the following optimization problem for all nonnegative values of the parameter λ:

$$\begin{gathered} \text{Minimize } V_{\mathrm{p}} - \lambda E_{\mathrm{p}} \\ \text{with respect to} X_1, X_2, \cdots, X_N \\ \text{subject to} \sum_i X_i = 1 \\ X_i \geqslant 0;\ i = 1, \cdots, N. \end{gathered} \tag{7.13}$$

The above problem is called the standard mean–variance portfolio selection problem. Nonstandard forms of the mean–variance portfolio selection problem include cases where: (1) additional linear constraints apply, and (2) the amounts invested in one or more securities may be negative. Markowitz (1956, 1959, 2000) describes efficient algorithms for solving both standard and nonstandard versions of the mean–variance portfolio selection problem. These algorithms are used extensively in practical mean–variance portfolio analysis.

Elton et al. (1976, 1978) demonstrate that an alternative simple procedure can be used to select mean–variance efficient portfolios when the single-index model (7.7) is accepted as the best method of forecasting mean–variance portfolio inputs. Their simple procedure requires that securities be ranked based on the ratio of their expected excess return to their beta:

$$\frac{(E_i - R_{\mathrm{f}})}{\beta_i}, \tag{7.14}$$

where E_i is the expected return on security i, R_{f} is the return on the risk-free security, and β_{i} is the sensitivity of the return on security i to changes in the market index as measured by (7.7). Elton, Gruber, and Padberg prove that all risky securities with an excess return to β ratio above a specific cut-off, C^*, should be included in a mean–variance efficient portfolio; and all risky securities with an excess return to β ratio below this cut-off value should be excluded. Formulas for C^* for both the case where short sales are not permitted and the case where short sales are permitted are given in Elton, Gruber, and Padberg's papers.

7.2.5 Choosing the "Best" Portfolio on the Efficient Frontier

In the Markowitz mean–variance model, an investor should always choose a portfolio on the mean–variance efficient frontier. However, the investor's choice of the best portfolio on the efficient frontier depends on his or her attitude toward risk. Risk-averse investors will likely choose efficient portfolios near the minimum risk portfolio, E, on the efficient frontier in Fig. 7.1 (or R_f on the efficient frontier in Fig.7.2), while risk-tolerant investors will likely choose portfolios near the maximum mean portfolio, F.

But how does the investor actually make the choice of the "best" portfolio on the efficient frontier? It appears that there are two alternatives: direct choice and investor utility functions. Direct choice involves a direct comparison of the mean and standard deviation of various portfolios on the efficient frontier. In the direct choice approach, the investor is simply presented with information on the means and standard deviations of various portfolios on the efficient frontier and asked to choose a preferred combination of mean and variance.

In contrast, investor utility functions involve an attempt to capture the investor's aversion to risk in the form of an investor-specific utility function:

$$U = E_p - kV_p, \tag{7.15}$$

where k is a parameter indicating the investor's aversion to risk, as measured by variance of return. In this approach, a portfolio advisor would estimate an investor's risk aversion parameter, k, from the investor's responses to a series of questions regarding the investor's attitude toward risk. Investors with high risk aversion would be assigned high values of k, while investors with low risk aversion would be assigned low values of k. The advisor would then calculate the utility, U, of each portfolio on the efficient frontier and recommend the portfolio with the highest utility.

In the tradition of Bernoulli (1738) and von Neumann and Morgenstern (1946), economists generally assume that investors wish to maximize their expected utility of wealth, an assumption that has allowed economists to derive a rich set of conclusions about investment behavior. However, practitioners have found that for most investors, utility functions are an impractical device for selecting portfolios. In their experience, they find that investors do not understand the concept of utility and are generally unable to provide the information required to determine their utility function analytically. Although this conclusion probably explains the minimal use of utility functions in practical portfolio analysis, it does not rule out using utility functions to obtain practical insights into optimal investment policies for typical investors. Indeed, we demonstrate below how utility analysis has produced many useful guidelines for lifetime consumption–investment decision making.

7.2.6 Comments on the Mean–Variance Model

The Markowitz mean–variance portfolio model has undoubtedly been one of the most influential models in the history of finance. Since its introduction in 1952, the mean–variance model has provided an intellectual foundation for much later research in finance. Because of its rich practical insights, the Markowitz model also continues to strongly influence practical financial management. Nonetheless, the mean–variance approach to portfolio selection is sometimes criticized because it implicitly assumes that information on a portfolio's mean return and variance of return is sufficient for investors to make rational portfolio decisions. Tobin (1958) notes that information on a portfolio's mean return and variance of return is only sufficient for rational portfolio decision making if: (1) the investor's utility function is quadratic; or (2) the probability distribution of security returns is normal. Neither of these assumptions is likely to be strictly true.

Economists generally agree that a reasonable utility function should display non-satiety, risk aversion, decreasing absolute risk aversion, and constant relative risk aversion.[6] The problem with a quadratic utility function is that it displays satiety (i.e., an investor with this utility function eventually prefers less wealth rather than more); and increasing absolute and relative risk aversion. Mossin (1968) demonstrates that the only utility functions that satisfy all four of the above desirable characteristics of utility functions are the logarithmic function, $\log W$,[7] and the power function, $W^{1-\gamma}$. Thus, the assumption that it is rational for investors to evaluate risky choices based solely on the mean and variance of return cannot be strictly justified on the grounds that investors' utility functions are quadratic.

The other way to justify the assumption that investors base their risky choices solely on the mean and variance of returns is that security returns are normally distributed. But this justification is also problematic. The normal distribution is symmetric with a positive probability that returns can take any value on the real line. However, with limited liability, an investor can never lose more than his entire wealth – that is, $(1 + r_t)$ must be greater than or equal to zero. In addition, the investor's multiperiod return is the product of individual period returns, and the product of normally distributed variables is not normally distributed. Thus the rationality of relying solely on the mean and variance of portfolio returns cannot be strictly justified on the grounds that returns are normally distributed.

Markowitz (1959), Levy and Markowitz (1979), and Samuelson (1970), among other prominent economists, recognize that investor utility functions are unlikely to be quadratic and that security return distributions are unlikely to be normally distributed. However, they defend the validity of the mean–variance approach to portfolio decision making based on the belief that one or more of the following statements is true:

- Within any interval, utility functions are approximately quadratic.
- Probability distributions can often be approximated by their first two moments.
- The mean–variance model, though not strictly rational, is nonetheless useful for investors because it provides information that investors find to be relevant and leads them to make better decisions than they would in the absence of the model.

[6] The desirable attributes of utility functions are discussed more fully below.

[7] In this chapter, we use the notation, $\log W$, to indicate the natural logarithm of W.

Indeed, Markowitz and Levy and Markowitz demonstrate that many utility functions are approximately quadratic in any interval; Samuelson demonstrates that probability distributions can under fairly general conditions be approximated by their first two moments; and the prevalence of the mean–variance framework in practical decision making suggests that the mean–variance model is useful to investors.

7.3 The Geometric Mean Portfolio Model

As described earlier, Markowitz achieves simplicity in the mean–variance model by focusing on the economic trade-off between risk and return in a single-period world. However, many investors make portfolio decisions in a multiperiod world where portfolios can be rebalanced periodically. For these investors, Latané (1959) recommends an alternative framework, the geometric mean portfolio model. He argues that the maximum geometric mean strategy will almost surely lead to greater wealth in the long run than any significantly different portfolio strategy, a result that follows from similar conclusions of Kelly (1957) in the context of information theory. Breiman (1960, 1961) states the precise conditions for which this result holds and develops additional properties of the geometric mean strategy. A recent news article describes how two well-known fund managers, Edward Thorp and Bill Gross, have used the geometric mean portfolio strategy to improve the performance of their funds.[8]

7.3.1 The Geometric Mean Strategy and Long-Run Wealth Maximization

Consider an investor who invests an amount, W_0, at the beginning of period 1 and earns a known rate of return on investment of $R_t = (1 + r_t)$ in periods $t = 1, \ldots, T$. If the investor reinvests all proceeds from his investment in each period, his wealth at the end of period T will be:

$$\begin{aligned} W_T &= W_0(1 + r_1)(1 + r_2) \cdots (1 + r_T) \\ &= W_0 \prod_t R_t. \end{aligned} \tag{7.16}$$

Let $G = (1 + g)$ denote the investor's compound average, or geometric mean, return on his investment over the period from 1 to T. Then,

$$\begin{aligned} G &= [(1 + r_1)(1 + r_2) \cdots (1 + r_T)]^{1/T} \\ &= \prod_t R_t^{1/T}. \end{aligned} \tag{7.17}$$

[8] "Old Pros Size Up the Game, Thorp and Pimco's Gross Open Up on Dangers of Over-betting, How to Play the Bond Market," *The Wall Street Journal*, Saturday/Sunday, March 22–23, 2008.

Since $W_T = W_0 G^T$, the investor's terminal wealth, W_T, will be maximized when the geometric mean return on investment, G, is maximized.

In practice, the returns $R_t = (1 + r_t)$ for $t = 1, \ldots, T$ are uncertain at the time the investor makes the initial investment decision. Assume that the return on investment is independently and identically distributed, and that there are J possible outcomes for R. Let P_j denote the probability of obtaining the jth return outcome. Then, the forward-looking geometric mean return on investment in this uncertain case is defined as:

$$G^E = \prod_j R_j^{P_j}. \tag{7.18}$$

In the geometric mean portfolio model, the investor's objective is to maximize the forward-looking geometric mean return on investment, G^E.

When analyzing a variable such as G or G^E that is equal to the product of other variables, it is frequently convenient to analyze the logarithm of the variable rather than the variable itself. From (7.18), the log of G^E is equal to the expected log return on investment:

$$\begin{aligned} \log G^E &= \sum_j P_j \log R_j \\ &= E \log R. \end{aligned} \tag{7.19}$$

Because the log function is monotonically increasing throughout its domain, any portfolio that maximizes G^E will also maximize $\log G^E$, and hence $E \log R$. Thus, the geometric mean portfolio strategy is equivalent to a portfolio strategy that seeks to maximize the expected log return on investment, $E \log R$.

Since the return on investment is assumed to be independently and identically distributed, the T values for R shown in the definition of G in (7.17) can be considered to be a random sample of size T from the probability distribution for R. Let G^S denote the geometric mean return calculated from the random sample of size T from the probability distribution for R and $\log G^S$ denote the log of the sample geometric mean return. From (7.17), $\log G^S$ is given by:

$$\log G^S = \frac{1}{T} \sum_t \log R_t. \tag{7.20}$$

According to the weak law of large numbers, the average of the sample values of a random variable will approach the expected value of the random variable as the sample size T approaches infinity. Presuming that the mean and variance of $\log R$ are finite, the weak law of large numbers assures that for any positive numbers ε and δ, no matter how small, there exists some positive number τ, perhaps large but nevertheless finite, such that for all $T > \tau$,

$$\text{Prob}\left\{ \left| \frac{1}{T} \sum_t \log R_t - E \log R \right| < \varepsilon \right\} \geqslant 1 - \delta. \tag{7.21}$$

An alternate notation for this condition is

$$\operatorname{plim}\left(\frac{1}{T}\sum_{t}\log R_t\right) = E\log R, \tag{7.22}$$

where plim denotes probability limit.

Let R^{A} denote the investor's returns under the maximum geometric mean strategy, and R^{B} denote the investor's returns under a significantly different strategy.[9] Then, from the above discussion, we know that:

$$\operatorname{plim}\left(\frac{1}{T}\sum_{t}\log R_t^{\mathrm{A}}\right) = E\log R^{\mathrm{A}}, \tag{7.23}$$

$$\operatorname{plim}\left(\frac{1}{T}\sum_{t}\log R_t^{\mathrm{B}}\right) = E\log R^{\mathrm{B}}, \tag{7.24}$$

and

$$E\log R^{\mathrm{A}} > E\log R^{\mathrm{B}}. \tag{7.25}$$

Thus,

$$\operatorname{plim}\left(\frac{1}{T}\sum_{t}\log R_t^{\mathrm{A}}\right) > \operatorname{plim}\left(\frac{1}{T}\sum_{t}\log R_t^{\mathrm{B}}\right). \tag{7.26}$$

This in turn implies that for T sufficiently large, it is virtually certain that

$$\left(\frac{1}{T}\right)\sum_{t}\log R_t^{\mathrm{A}}$$

will exceed

$$\left(\frac{1}{T}\right)\sum_{t}\log R_t^{\mathrm{B}},$$

and do so by an amount very nearly equal to $(E\log R^{\mathrm{A}} - E\log R^{\mathrm{B}}) > 0$. That is, in the long run, the investor will almost surely have greater wealth by using the geometric mean strategy than by using any significantly different strategy.

7.3.2 The Relationship Between the Geometric Mean Strategy and the Log Utility Function

As shown above, the geometric mean portfolio strategy: (1) almost surely produces greater wealth in the long run than any significantly different strategy; and (2) is

[9] By "significantly different strategy," we mean a strategy that has a lower expected log return than the maximum geometric mean strategy.

equivalent to a strategy that seeks to maximize the expected log return. Further, Mossin (1968) demonstrates that for the log and power utility functions, maximizing the expected utility of return is equivalent to maximizing the expected utility of terminal wealth. Thus, for investors with log utility functions, the maximum geometric mean portfolio criterion is equivalent to the maximum expected utility of wealth criterion.

7.3.3 Desirable Properties of the Log Utility Function

For investors with log utility functions, the equivalence of the maximum geometric mean and the maximum expected log utility criteria is significant because log utility functions have many desirable properties. Among these properties are: (1) non-satiety; (2) risk aversion; (3) decreasing absolute risk aversion; (4) constant relative risk aversion; (5) aversion to margin investing (i.e., investing with borrowed money); and (6) optimality of myopic decision making. Mossin (1968) demonstrates that the log and power functions are the only utility functions with all of these properties.

Non-satiety. Because the log utility function is increasing throughout its domain, investors with log utility functions always prefer more wealth to less, an attribute that economists refer to as "non-satiety." Although non-satiety would seem to be an obvious desirable property of a utility function, many utility functions do not possess this property, including the quadratic utility function.

Risk aversion. Consider the choice between (1) receiving \$W for certain and (2) having a 50/50 chance of receiving either \$$(W + C)$ or \$$(W - C)$. Investors who choose alternative (1) over alternative (2) are said to be risk averse, because these alternatives have the same expected value, but the second alternative has greater risk. Risk-averse investors choose alternative (1) because the difference in utility from receiving \$$(W + C)$ rather than \$$(W)$ is less than the difference in utility from receiving \$ $(W - C)$ rather than \$ (W). That is, $U(W + C) - U(W) < U(W) - U(W - C)$. Evidently, risk-averse investors have utility functions characterized by $U''(W) < 0$.[10] Since the second derivative of log W is $-1/W^2$, investors with log utility functions are risk averse.

Decreasing absolute risk aversion. Although all risk-averse investors have utility functions characterized by $U''(W) < 0$, some investors are more risk averse than others. The intensity of an investor's aversion to risk is determined by the curvature of the investor's utility function, where curvature is measured by the ratio of $U''(W)$ to $U'(W)$. Specifically, Pratt (1964) and Arrow (1965) define the coefficient of absolute risk aversion by the equation:

$$\mathrm{ARA}(W) = -\frac{U''(W)}{U'(W)}. \tag{7.27}$$

[10] We use the notation $U''(W)$ to indicate the second derivative of the utility function U with respect to its argument, W.

For small gambles, Pratt and Arrow demonstrate that $\text{ARA}(W)$ determines the dollar amount an investor is willing to pay to avoid a fair gamble with the possibility of either winning or losing a constant absolute dollar amount, C. They argue that absolute risk aversion should decrease with wealth because rich investors would be relatively unconcerned with losing an amount, C, that would cause great concern for poor investors. For the log utility function, $\text{ARA}(W)$ equals $1/W$. Thus, log utility functions imply that absolute risk aversion declines as wealth increases.

Constant relative risk aversion. Pratt and Arrow define the coefficient of relative risk aversion by the equation:

$$\text{RRA}(W) = -\frac{WU''(W)}{U'(W)}. \tag{7.28}$$

For small gambles, they demonstrate that $\text{RRA}(W)$ determines the fraction of wealth that an investor will pay to avoid a fair gamble with a possibility of winning or losing a specific fraction, C/W, of wealth. Economists generally believe that relative risk aversion should remain constant as wealth increases. This belief is consistent with the evidence that interest rates and risk premia have remained constant as average wealth has increased over time. Since $\text{RRA}(W)$ equals 1 for the log utility function, log utility functions display constant relative risk aversion.

Aversion to investing on margin. Investors can frequently enhance their expected return on investment by investing on margin. However, margin investing is considered to be risky for individual investors because it can greatly increase both the variability of return on investment and the probability of bankruptcy. Investors with log utility functions are averse to margin investing because margin investing increases the probability that wealth will be less than some small value ε greater than zero; and their utility of wealth approaches minus ∞ as W approaches zero.

Optimality of myopic decision making. As noted above, lifetime portfolio selection is generally a complex problem that, because of its dynamic interdependence, can only be solved through sophisticated dynamic programming procedures. However, Mossin (1968), Samuelson (1969), and Merton (1969) demonstrate that investors with either log or power utility functions can solve their lifetime portfolio selection problem one period at a time. For these investors, the decision that maximizes the expected utility of wealth at the end of period t is the same as the tth period decision resulting from a dynamic optimization procedure that considers the effect of the individual's tth period decision on all future decisions. Myopic decision making is a highly desirable property of utility functions because it allows analytical solutions to problems that would otherwise be impossible to solve.

7.3.4 Solving the Geometric Mean Portfolio Problem

An optimal geometric mean portfolio consists of a set of securities, $i = 1, \ldots N$, and the optimal proportion of wealth, X_i, to invest in each security. Let $R_i = (1 + r_i)$

denote the return on security i. Then the optimal geometric mean portfolio can be found by solving the following optimization problem:

$$\begin{gathered}\text{Maximize } E\left[\log \sum_{i=1}^{N} R_i X_i\right] \\ \text{with respect to} X_1, \cdots, X_N \\ \text{subject to} X_i \geqslant 0,\ i = 1, \cdots, N \\ \sum_{i=1}^{N} X_i = 1.\end{gathered} \tag{7.29}$$

Vander Weide et al. (1977) establish conditions required for the existence of a solution to the geometric mean portfolio problem and provide computational methods for finding exact solutions when solutions do exist. Maier et al. (1977) examine the solutions to a relatively large number of simulated geometric mean portfolio problems obtained with the aid of a numerically efficient nonlinear programming code embedded within a partitioning algorithm. They find that the number of risky securities in an optimal geometric mean portfolio depends on one's expectations concerning future market conditions and on the conditions under which borrowing is permitted. When borrowing is not permitted, the investor who believes the market will fall should invest in just one security; the investor who believes the market will remain unchanged should diversify among two securities; and the investor who believes the market will rise should diversify among four to seven securities.[11]

When borrowing is allowed, the geometric mean portfolio problem must be modified to assure that the investor will not go bankrupt. Avoidance of bankruptcy can be accomplished by requiring the investor to withhold sufficient capital from investment to cover interest and principal payments on the borrowing. In the modified geometric mean portfolio problem, the investor who believes the market will rise should choose the same securities in the same relative proportions as when no borrowing is allowed. Furthermore, the individual characteristics of securities contained in optimal geometric mean portfolios also depend on one's assumptions about market conditions and the availability of borrowing. If a rising market is anticipated, the investor should invest in securities for which β_i and σ_i are large, and for which α_i is small. If the market is expected to decline, the investor should invest in stocks with high α_i and low β_i and σ_i.

Maier et al. (1977) also describe several heuristic portfolio building rules that provide near-optimal solutions to the geometric mean portfolio problem, including a geometric mean rule, a reward-to-variability rule, a reward-to-nondiversifiable variability rule, and a Kuhn–Tucker rule. Each rule follows the same principle: rank each security on the basis of one criterion, and then allocate equal dollar amounts

[11] These numbers of securities are obtained under the assumption that the investor has 100 securities from which to choose. If the investor can choose among a greater number of securities, the optimal number to hold is likely to increase.

to several of the top-ranked securities. They find that the geometric mean rule, the reward-to-variability rule, and the Kuhn-Tucker rule provide reasonable approximations to the returns on the optimal geometric mean portfolio.

Of course, the geometric mean portfolio problem cannot be solved without appropriate data inputs. Maier et al. (1977) note that the data inputs to the geometric mean portfolio problem can be estimated by assuming that the distribution of the holding period return, R_i, is related to a market index, I, through the equation:

$$\log R_i = \alpha_i + \beta_i I + \varepsilon_i, \tag{7.30}$$

where I is defined as the expected value over all securities of the logarithm of the holding period return, α_i and β_i are constants, and ε_i is a normal random variable with mean zero and variance σ_i^2. The index I is considered a normal random variable whose parameters are chosen subjectively by the investor. Maier et al. (1982) develop an empirical Bayes estimation procedure for obtaining a simultaneous estimate of the three market model parameters of (7.30) that makes use of more information than other estimates described in the literature.

7.3.5 *Relationship Between the Maximum Geometric Mean Return and the Mean and Variance of Return on a Portfolio*

The geometric mean portfolio strategy is specifically designed for investors who wish to maximize their long-run wealth. Since the "long run" may be many years in the future, however, and mean–variance efficient portfolios have desirable short-run properties, it is natural to inquire whether the maximum geometric mean portfolio is mean–variance efficient. Markowitz (1959) and Young and Trent (1969) address this inquiry by examining the expected value of several Taylor series approximations of either $E \log R$ or G. In discussing their methods and results, we will use the following notation:

$\mu_1 = E(R) =$ the expected value or first moment of the probability distribution of R
$\mu_2 = E(R-\mu_1)^2 =$ the variance or second central moment
$\mu_3 = E(R-\mu_1)^3 =$ the skewness or third central moment
$\mu_4 = E(R-\mu_1)^4 =$ the kurtosis or fourth central moment

The Taylor series approximation of log R centered around the mean value, μ_1, of R is given by:

$$\log R = \log \mu_1 + \frac{R-\mu_1}{\mu_1} - \frac{(R-\mu_1)^2}{2\mu_1^2} + \frac{(R-\mu_1)^3}{3\mu_1^3} - \frac{(R-\mu_1)^4}{4\mu_1^4} + \cdots \tag{7.31}$$

Taking expectations of both sides of (7.31), and noting that $E(R - \mu_1) = 0$, we then have:

$$E \ \log R = \log \mu_1 - \frac{\mu_2}{2\mu_1^2} + \frac{\mu_3}{3\mu_1^3} - \frac{\mu_4}{4\mu_1^4} + \cdots \tag{7.32}$$

Equation (7.32) provides several important insights about the relationship between the maximum geometric mean return and the mean and variance of return on a portfolio. First, if the third and higher moments of the probability distribution of R are "small" in relation to the first moment, $E \log R$ can be reasonably approximated by the expression:

$$E \ \log R = \log \mu_1 - \frac{\mu_2}{2\mu_1^2}. \tag{7.33}$$

Second, if (7.33) is a reasonable approximation for $E \log R$, the geometric mean portfolio will be approximately mean–variance efficient because $E \log R$ will be maximized when the mean, μ_1, is maximized for any value of variance, μ_2, and the variance, μ_2, is minimized for any value of mean, μ_1. Third, if the third and higher moments of the probability distribution for R are not "small" in relation to the first moment, the geometric mean portfolio may not be mean–variance efficient. An example where the geometric mean portfolio is not mean–variance efficient is provided by Hakansson (1971).

To test whether the maximum geometric mean portfolio is approximately mean variance efficient, Markowitz (1959) examines the ability of two geometric mean approximations to $E \log R$ to predict the actual geometric mean return on nine securities and two portfolios over the period 1937–1954. The two geometric mean approximations include:[12]

$$G\,(1) \ = \mu_1 - \frac{\mu_1^2 + \mu_2^2}{2},$$

and

$$G\,(2) \ = \log\mu_1 - \frac{\mu_2}{2\mu_1^2}.$$

He finds that $G(1)$ consistently underestimates the geometric mean return on the nine securities and two portfolios over the period, with an average error of 8%. However, $G(2)$ performs significantly better than $G(1)$. It slightly overestimates the actual geometric mean return with an average error of only 1.7%. From his analysis of these approximations, Markowitz suggests that $G(2)$ be used to estimate the geometric mean return for each portfolio on the mean–variance efficient frontier. He advises investors never to choose portfolios on the mean–variance efficient frontier

[12] $G(1)$ is derived from a Taylor series approximation centered on $R = 1$, while $G(2)$ is the approximation shown in (7.33).

with greater single-period means than the optimal geometric mean portfolio because such portfolios will have higher short-run variance than the optimal geometric mean portfolio and less wealth in the long run.

Using methods similar to Markowitz, Young and Trent (1969) empirically test the ability of five geometric mean approximations to predict the actual geometric mean return on 233 securities and various portfolios based on these securities. The five geometric mean approximations include:[13]

$$G\,(1) \;= (\mu_1^2 - \mu_2)^{1/2},$$

$$G\,(2) \;= \mu_1 - \frac{\mu_2}{2\mu_1},$$

$$G\,(3) \;= \mu_1 - \frac{\mu_2}{2},$$

$$G\,(4) \;= \mu_1 - \frac{\mu_2}{2\mu_1} + \frac{\mu_3}{3\mu_1^2},$$

and

$$G\,(5) \;= \mu_1 - \frac{\mu_2}{2\mu_1} + \frac{\mu_3}{3\mu_1^2} - \frac{\mu_4}{4\mu_1^3}.$$

Using monthly holding period returns for the time period January 1957 to December 1960 and annual holding period returns for the period January 1953 to December 1960, they demonstrate that geometric mean approximations such as $G(2)$ and $G(3)$, based only on the mean and variance of the probability distribution of R, provide predictions of geometric mean returns that differ on average from actual geometric mean returns by 0.5%. Thus, for their data set, we may conclude that maximum geometric mean portfolios are highly likely to be mean–variance efficient.

7.3.6 Comments on the Geometric Mean Portfolio Model

The geometric mean portfolio strategy is designed for investors who seek to maximize the expected value of their wealth in the long run. However, Merton and Samuelson (1974) demonstrate that maximizing the expected value of long-run wealth is not the same as maximizing the expected utility of long-run wealth. Since wealth at the end of a typical lifetime is variable, investors who are more risk averse than investors with log utility functions may prefer an alternative investment strategy that provides a stronger hedge against values of wealth that are less than the expected value of long-run wealth.

[13] These approximations are derived from the Taylor series expansion of G centered around the mean value of R.

7.4 Lifetime Consumption–Investment Model

The mean–variance and geometric mean portfolio models are designed to help investors choose the optimal proportions of wealth to invest in each security, based only on information regarding the probability distributions of returns on securities. If the probability distributions of returns are assumed to be independently and identically distributed, these models will recommend that the proportion of wealth invested in each security remain constant over the investor's lifetime. However, a constant proportion investment strategy is inconsistent with conventional wisdom that investors, as they age, should lower the proportion of wealth invested in risky stocks vs. less risky bonds. The lifetime consumption–investment model is designed to help investors understand the conditions under which their optimal investment policy might change over their lifetimes, even if their probability beliefs remain constant.

Interest in lifetime consumption–investment models began in the late 1960s. Important early papers include Samuelson (1969), Merton (1969), Mossin (1968), and Hakansson (1969, 1970). Important later papers include Viceira (2001), Heaton and Lucas (2000), Koo (1998, 1999), Campbell et al. (2001a, b), Bodie et al. (1992), and Campbell and Cochrane (1999). Campbell and Viceira (2002) contains an excellent discussion of lifetime consumption–investment models, as well as a review of the literature on this important topic.

7.4.1 *The Standard Lifetime Consumption–Investment Model*

Consider an individual who must choose the amounts to consume, C_t, the fraction of wealth to invest in risky assets, w_t, and the fraction of wealth to invest in a risk-free asset, $(1-w_t)$, at the beginning of each period ($t = 0, 1, \ldots, T$). Assume that the individual's goal is to maximize the expected present value of the utility from lifetime consumption and that wealth must be either consumed or invested in each period. Let Z_t denote the random return on the risky asset, ρ the investor's discount rate, and r, the return on the risk-free asset. Then, the individual's standard lifetime consumption–investment problem can be stated as:[14]

$$\begin{gathered}\text{Max } E\left[\sum_{t=0}^{T}(1+\rho)^{-t}U(C_t)\right]\\ \text{with respect to } C_t, w_t\\ \text{subject to } C_t = \left[W_t - \frac{W_{t+1}}{(1+r)(1-w_t)+w_t Z_t}\right]\\ W_0 \text{ given, } W_{T+1} \text{ prescribed.}\end{gathered} \tag{7.34}$$

[14] This formulation is taken from Samuelson (1969).

7.4.2 Analysis of the Optimal Lifetime Consumption–Investment Strategy

The standard formulation of the lifetime consumption–investment problem is difficult to solve without some simplifying assumptions. In his first paper on this subject, Samuelson (1969) assumes that (1) the individual's utility function displays constant relative risk aversion, that is, the utility function is either a log or power function; and (2) the probability distribution for Z_t is independently and identically distributed. To his surprise, he finds that the optimal proportion to invest in the risky asset is constant under these assumptions. Thus, under the standard assumptions, the lifetime consumption–investment model produces the same constant proportion recommendation as the mean–variance and geometric mean models. Merton, Leland, Mossin, and Hakansson reach similar conclusions.

However, Samuelson (1989), Bodie et al. (1992), and the authors of later papers cited above demonstrate that when the standard assumptions of the lifetime consumption–investment model are modified to include nontradable human wealth,[15] subsistence levels of consumption, and mean-reverting probability distributions of returns, the conclusion that the percentage invested in risky assets is constant must be modified. Since the literature on the effect of these additional variables on the optimal solution to the lifetime consumption–investment problem is complex, we limit our discussion here to a brief summary of relevant conclusions.

Nontradable human wealth. The effect of human wealth on an individual's optimal investment strategy depends on whether human wealth is riskless or risky. Assume first that human wealth is riskless, that is, that the present value of an individual's future income is certain. Since riskless human wealth is equivalent to an implicit investment in a riskless asset, the investor should adjust the proportion of financial wealth, F_t, invested in risky assets to reflect the investor's implicit additional holding of riskless assets. When human wealth is riskless, Campbell and Viceira demonstrate that the optimal proportion of financial wealth, F_t, to invest in risky assets is an increasing function of the ratio of human wealth to financial wealth, H_t/F_t. This ratio will typically vary over an individual's lifetime.

For young investors, the ratio of human wealth to financial wealth will typically be high because the young investor (1) can expect to earn labor income for many years to come; and (2) has not had much time to accumulate financial wealth. Thus, young investors should allocate a relatively large percentage of financial wealth to risky assets. In contrast, for investors nearing retirement, the ratio of human wealth to financial wealth will typically be low. Thus, when human wealth is riskless, the percentage of financial wealth invested in risky assets should decline with age.

[15] Human wealth, H_t, reflects the expected present value of an individual's future income. Human wealth is nontradable because the legal system forbids trading in claims on an individual's future income. Financial wealth, F_t, reflects the current market value of an individual's financial assets, that is, stocks and bonds. Total wealth, W_t, is equal to H_t plus F_t.

Assume now that labor income, and hence human wealth, is risky. If labor income is uncorrelated with the return on risky assets, the investor with human wealth should still invest a greater percentage of financial wealth in risky assets than the investor without human wealth. However, the percentage invested in risky assets should decrease with increases in the variance of labor income, that is, investors with high variance in labor income should reduce the percentage of financial wealth invested in risky assets to hedge some of the risk of their labor income.

If labor income is perfectly positively correlated with the return on one or more risky financial assets, human wealth is an implicit investment in these financial assets. In this case, the investor should either increase the percentage of financial wealth invested in the riskless asset or diversify into risky assets that are uncorrelated with labor income. This latter conclusion applies to all individuals who hold a high percentage of financial wealth in the stock of their employer.

Human wealth also affects the optimal percentage of financial wealth to invest in risky assets through the investor's ability to vary his or her work effort. If the investor can increase work effort to offset losses on financial assets, the optimal percentage of financial wealth to invest in risky assets will increase.

Subsistence levels of consumption. The optimal percentage of financial wealth to invest in risky assets also depends on the investor's desire to maintain a minimum level of consumption. A minimum level of consumption may be thought of as negative income because a certain part of income must be set aside to assure the minimum level of consumption. Samuelson (1989) establishes that as the investor nears retirement, the investor should increase the allocation of financial wealth to risk-free bonds to assure a minimum level of consumption in retirement. The shift toward risk-free bonds arises as a result of the investor's need to provide a steady income stream in retirement to cover the minimum level of consumption. However, Merton (1969) notes that young investors may also have a need to assure a minimum level of consumption in the face of uncertain human and financial wealth. He establishes that this need would also shift the optimal portfolio of young investors toward riskless bonds. Constantinides (1990) and Campbell and Cochrane (1999) analyze the case where the minimum level of consumption itself may depend on either the individual's prior consumption habits or the consumption norms of society.

Mean-reverting probability distribution of returns. Samuelson (1989) demonstrates that when asset returns are mean reverting, investors with long investment horizons should invest more in risky assets than investors with short investment horizons. Campbell, Cocco, Gomes, Maenhout, and Viceira (2001) show that if asset returns are mean reverting, investors should reduce the percentage of financial wealth invested in risky assets when the returns on risky assets have recently been above the long-run mean return and increase the percentage of financial wealth invested in risky assets when returns on risky assets have recently been below the long-run mean return. Thus, investors who believe that returns on risky assets are mean reverting should also vary the percentage of financial wealth invested in risky assets with the status of the capital markets.

Acknowledgment I appreciate the helpful comments of John Guerard, David Peterson, David Vander Weide, and two anonymous referees.

References

Arrow, Kenneth J., 1965, Aspects of the theory of risk-bearing (Yrjö Jahnssonin Säätiö, Helsinki, Finland).

Bernoulli, Daniel, 1738, Exposition of a new theory on the measurement of risk, Papers of the Imperial Academy of Sciences in Petersburg V, 175–192.

Bodie, Zvi, Robert C. Merton, and William F. Samuelson, 1992, Labor supply flexibility and portfolio choice in a life cycle model, Journal of Economic Dynamics and Control 16, 427–449.

Breiman, Leo, 1960, Investment policies for expanding businesses optimal in a long-run sense, Naval Research Logistics Quarterly 7, 647–651.

Breiman, Leo, 1961, Optimal gambling systems for favorable games, in Jerzy Neyman, ed.: *Proceedings of the Fourth Berkeley Symposium on Mathematics and Probability* (University of California Press, California).

Campbell, John Y., João Cocco, Francisco Gomes, and Pascal Maenhout, 2001a, Investing retirement wealth: a life-cycle model, in John Y. Campbell and Martin Feldstein, ed.: *Risk Aspects of Investment-based Social Security Reform* (University of Chicago Press, Chicago).

Campbell, John Y., João Cocco, Francisco Gomes, and Pascal Maenhout, 2001b, Stock market mean reversion and the optimal equity allocation of a long-lived investor, *European Finance Review* 5.

Campbell, John Y. and John H. Cochrane, 1999, By force of habit: a consumption-based explanation of aggregate stock market behavior, Journal of Political Economy 107, 205–251.

Campbell, John Y., and Luis M. Viceira, 2002, *Strategic Asset Allocation: Portfolio Choice for Long-term Investors* (Oxford University Press, New York).

Campbell, John Y., João Cocco, Francisco Gomes, and Pascal Maenhout, 2001, Stock market mean reversion and the optimal equity allocation of a long-lived investor, *European Finance Review* 5.

Chapman, S. J., 2006, The Kelly criterion for spread bets, IMA Journal of Applied Mathematics 72, 43–51.

Constantinides, George M., 1990, Habit formation: A resolution of the equity premium puzzle, Journal of Political Economy 98, 519–543.

Elton, Edwin J., and Martin J. Gruber, 1974a, On the maximization of the geometric mean with lognormal return distribution, Management Science 21, 483–488.

Elton, Edwin J., and Martin J. Gruber, 1974b, On the optimality of some multiperiod portfolio selection criteria, The Journal of Business 47, 231–243.

Elton, Edwin J., Martin J. Gruber, and Manfred W. Padberg, 1976, Simple criteria for optimal portfolio selection, The Journal of Finance 31, 1341–1357.

Elton, Edwin J., Martin J. Gruber, and Manfred W. Padberg, 1978, Simple criteria for optimal portfolio selection: tracing out the efficient frontier, The Journal of Finance 33, 296–302.

Fabozzi, Frank J., Francis Gupta, and Harry M. Markowitz, 2002, The legacy of modern portfolio theory, Journal of Investing 11, 7.

Fama, Eugene F., 1970, Multiperiod consumption-investment decisions, The American Economic Review 60, 163–174.

Grauer, Robert R., 1981, A comparison of growth optimal and mean variance investment policies, The Journal of Financial and Quantitative Analysis 16, 1–21.

Hakansson, Nils H., 1969, Optimal investment and consumption strategies under risk, an uncertain lifetime, and insurance, International Economic Review 10, 443–466.

Hakansson, Nils H., 1970, Optimal investment and consumption strategies under risk for a class of utility functions, Econometrica 38, 587–607.

Hakansson, Nils H., 1971, Capital growth and the mean-variance approach to portfolio selection, The Journal of Financial and Quantitative Analysis 6, 517–557.
Hakansson, Nils H., and Tien-Ching Liu, 1970, Optimal growth portfolios when yields are serially correlated, The Review of Economics and Statistics 52, 385–394.
Heaton, John, and Deborah J. Lucas, 2000, Portfolio choice and asset prices: the importance of entrepeneurial risk, *Journal of Finance* 55, 1163–1198.
Jacobs, Bruce I., Kenneth N. Levy, and Harry M. Markowitz, 2005, Portfolio optimization with factors, scenarios, and realistic short positions, Operations Research 53, 586–599.
Jacobs, Bruce I., Kenneth N. Levy, and Harry M. Markowitz, 2006, Trimability and fast optimization of long-short portfolios, Financial Analysts Journal 62, 36–46.
Jin, Hanqing, Harry Markowitz, and Zhou Xun Yu, 2006, A note on semivariance, Mathematical Finance 16, 53–61.
Kelly, J. L., 1956, A new interpretation of information rate, Bell Systems Technical Journal 35, 917–926.
Koo, Hyeng Keun, 1998, Consumption and portfolio selection with labor income: a continuous time approach, *Mathematical Finance* 8, 49–65.
Koo, Hyeng Keun, 1999, Consumption and portfolio selection with labor income: a discrete-time approach, *Mathematical Methods of Operations research* 50, 219–243.
Kroll, Yoram, Haim Levy, and Harry M. Markowitz, 1984, Mean-variance versus direct utility maximization, The Journal of Finance 39, 47–61.
Latané, Henry Allen, 1959, Criteria for choice among risky ventures, The Journal of Political Economy 67, 144–155.
Latané, Henry Allen, 1960, Individual Risk Preference in Portfolio Selection, The Journal of Finance 15, 45–52.
Latané, Henry A., 1963, Investment criteria: a three asset portfolio balance model, The Review of Economics and Statistics 45, 427–430.
Latané, Henry A., 1979, The geometric mean criterion continued, Journal of Banking and Finance 3, 309–311.
Latané, Henry A., and Donald L. Tuttle, 1966, Decision theory and financial management, The Journal of Finance 21, 228–244.
Latané, Henry A., and Donald L. Tuttle, 1967, Criteria for portfolio building, The Journal of Finance 22, 359–373.
Latané, Henry A., and William E. Young, 1969, Test of portfolio building rules, The Journal of Finance 24, 595–612.
Levy, H., and H. M. Markowitz, 1979, Approximating expected utility by a function of mean and variance, The American Economic Review 69, 308–317.
Maier, Steven F., David W. Peterson, and James H. Vander Weide, 1977, A Monte Carlo investigation of characteristics of optimal geometric mean portfolios, The Journal of Financial and Quantitative Analysis 12, 215–233.
Maier, Steven F., David W. Peterson, and James H. Vander Weide, 1982, An empirical Bayes estimate of market risk, Management Science 28, 728–737.
Markowitz, Harry M., 1952, Portfolio selection, The Journal of Finance 7, 77–91.
Markowitz, Harry M., 1959, Portfolio selection: efficient diversification of investments (Yale University Press, New Haven).
Markowitz, Harry M., 1976, Investment for the long run: new evidence for an old rule, The Journal of Finance XXXI, 1273–1286.
Markowitz, Harry M., 1987. Mean-variance analysis in portfolio choice and capital markets (B. Blackwell, New York).
Markowitz, Harry M., 1991. Portfolio selection: efficient diversification of investments (B. Blackwell, Cambridge, Mass.).
Markowitz, Harry M., 1999a, The early history of portfolio theory: 1600 - 1960, Financial Analysts Journal 5–16.
Markowitz, Harry M., 1999b, Normative portfolio analysis: past, present, and future, Journal of Economy and Business 42, 99–103.

Markowitz, Harry M., 2000. Mean-variance Analysis in Portfolio Choice and Capital Markets (J. W. Wiley, New York).

Markowitz, Harry M., 2002, Efficient portfolios, sparse matrices, and entities: a retrospective, Operations Research 50, 154–160.

Markowitz, Harry M., 2005, Market efficiency: a theoretical distinction and so what?, Financial Analysts Journal 61, 17–30.

Markowitz, Harry M., and Erik L. van Dijk, 2003, Single-period mean-variance analysis in a changing world, Financial Analysts Journal 59, 30–44.

Markowitz, Harry M., and Andre F. Perold, 1981, Portfolio analysis with factors and scenarios, The Journal of Finance XXXVI, 871–877.

Merton, Robert C., 1969, Lifetime portfolio selection under uncertainty: the continuous-time case, *Review of Economics & Statistics* LI.

Merton, Robert C., and Paul A. Samuelson, 1974, Fallacy of the log-normal approximation to optimal portfolio decision-making over many periods, Journal of Financial Economics 1, 67–94.

Mossin, Jan, 1968, Optimal multiperiod portfolio policies, The Journal of Business 41, 215–229.

Pratt, J., 1964, Risk aversion in the small and in the large, Econometrica 32.

Roy, A. D., 1952, Safety first and the holding of assets, Econometrica 20, 431–499.

Samuelson, Paul A., 1969, Lifetime portfolio selection by dynamic stochastic programming, Review of Economics & Statistics 51, 239.

Samuelson, Paul 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.

Samuelson, Paul A., 1979, Why we should not make mean log of wealth big though years to act are long, Journal of Banking & Finance 3, 305–307.

Samuelson, Paul A., 1989, A case at last for age-phased reduction in equity, Proceedings of the National Academy of Sciences of the United States of America 86, 9048–9051.

Sharpe, William F., 1963, A simplified model for portfolio analysis, Management Science 9, 277–293.

Sharpe, William F., 1967, Portfolio analysis, The Journal of Financial and Quantitative Analysis 2, 76–84.

Tobin, James, 1958, Liquidity preference as behavior towards risk, The Review of Economic Studies XXVI, 65–86.

Vander Weide, James H., David W. Peterson, and Steven F. Maier, 1977, A strategy which maximizes the geometric mean return on portfolio investments, Management Science 23, 1117–1123.

Vander Weide, James H., David W. Peterson, and Steven F. Maier, 1978, A comment on geometric mean portfolios: reply to Aucamp, Management Science 24, 860.

Viceira, Luis M., 2001, Optimal portfolio choice for long-horizon investors with nontradable labor income, *Journal of Finance* 56, 433–470.

von Neumann John, and Oskar Morgenstern, 1946. Theory of Games and Economic Behavior (Wiley, New York).

Young, William E., and Robert H. Trent, 1969, Geometric mean approximations of individual securities and portfolio performance, Journal of Financial & Quantitative Analysis 4, 179.

Ziemba, William T., 1972, Note on 'optimal growth portfolios when yields are serially correlated', Journal of Financial & Quantitative Analysis 7, 1995.

Chapter 8
Harry Markowitz and the Early History of Quadratic Programming

Richard W. Cottle and Gerd Infanger

8.1 Introduction

Harry M. Markowitz, 1990 Nobel Laureate in Economics, is widely known in optimization and investment science circles as the principal architect of the portfolio selection problem. Apart from its practical value in the world of investment, the portfolio selection problem articulated by Markowitz represents one of the most frequently identified and discussed examples of a quadratic programming problem, though not always as the *parametric* quadratic programming problem that it actually is. Quadratic programs more frequently arise as subproblems to be solved in the process of solving some other problem; thus when compared with most applications of quadratic programming, the portfolio selection problem looks quite "natural." It is often presented as if the solution to the portfolio selection problem were precisely the information being sought. This may account for its great appeal to the authors of papers and books on quadratic and nonlinear programming.

But important and often discussed as Markowitz's *model* for the portfolio selection problem may be, his *quadratic programming algorithm*, the Critical Line Algorithm (CLA) he devised in Markowitz (1956) to solve such problems, has – after more than 50 years – received rather little attention in the literature of mathematical programming. Writing in the Foreword of the 2000 edition of *Mean–Variance Analysis in Portfolio Choice and Capital Markets* (Markowitz (2000)), William F. Sharpe succinctly captures the state of affairs saying:

> Markowitz's early works have suffered the fate of those of other pioneers: often cited, less often read (at least completely). Indeed, in this book he evidences his concern that "many scholars interested in such matters apparently never found [the discussion of fundamentals in the back of the 1959 book.]"

We document this observation using the small body of contemporaneous quadratic programming literature.

Perhaps the first to cite Markowitz's quadratic programming paper was Philip Wolfe who, in his paper on a simplex method for quadratic programming (Wolfe

---

R.W. Cottle (✉) and G. Infanger
Department of Management Science and Engineering, Stanford University, Stanford, CA 94305–4026, USA
e-mail: rwc@stanford.edu; infanger@stanford.edu

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_8,
© Springer Science+Business Media, LLC 2010

(1959)), wrote "Markowitz ... has suggested a method ... for solving the 'portfolio' problem ... The method described here exploits this ingenious idea, differing from the proposal of Markowitz (1956) in keeping to a linear programming format." Wolfe op. cit. says nothing more about Markowitz's algorithm. Indeed, he extends as much credit to Barankin and Dorfman (1958) as to Markowitz (1956). The brief abstract of Wolfe's paper includes the remark that "the procedure is analogous to the simplex method for linear programming, being based on the Barankin-Dorfman procedure for this problem." And later, in the body of the article (Wolfe, 1959, p. 283), Wolfe says "Barankin and Dorfman ... first pointed out the linear formulation of the problem, inspiring the present approach." Another of Wolfe's papers (Wolfe (1963)) is a survey on methods of nonlinear programming containing a short section (little more than one page) on simplex methods for quadratic programming. The discussion touches on only two papers: Beale (1959) and Wolfe (1959), both of which date from the same year. The paper makes no mention of Markowitz or his paper of 1956. For that matter, neither does Beale's paper.

Markowitz's RAND report Markowitz (1955), the precursor to the paper Markowitz (1956), is listed among the references in Jack Bonnell Dennis's book Dennis (1959), but it is not actually cited there. Samuel Karlin's book (Karlin (1959)) published in the same year describes the portfolio selection problem and cites the source. Karlin, like Dennis, lists the aforementioned RAND report, but does not mention it in the text and does not describe the CLA.

The paper Theil and van de Panne (1960) addresses the prior publications Beale (1955), Beale (1959), Hildreth (1957), Houthakker (1959), Markowitz (1959), and Wolfe (1959) in a footnote saying "For other contributions to the problem of quadratic programming, see the list of references at the end of this paper."

In (Zoutendijk, 1960, p. 110), Guus Zoutendijk's book/doctoral dissertation *Methods of Feasible Directions*, the author cites Markowitz's algorithm (Markowitz (1956)) among "a number of well-known quadratic programming methods." Zoutendijk then briefly discusses the work of Beale, Frank and Wolfe, and Wolfe, but not that of Markowitz.

Later George Dantzig released a technical report (Dantzig (1961)), which he called "Quadratic Programming: A variant of the Wolfe-Markowitz algorithms".[1] After acknowledging the work of Markowitz, Dantzig says "We shall present here a variant of Wolfe's procedure ..." implicitly suggesting that the algorithms of Markowitz and Wolfe are different. As stated, Dantzig's algorithm is a variant of Wolfe's "short form" algorithm for minimizing a positive semidefinite quadratic form subject to linear equations in nonnegative variables. Dantzig does not address the parametric aspect of the problem.

The German edition of Künzi and Krelle (1966) published in 1962 (and later translated into English) opens a chapter called "Wolfe's Method" with the statement "Following a suggestion made by Markowitz (Markowitz (1956)), Wolfe (Wolfe (1959)) has developed a method for solving quadratic programs, which uses the

[1] This material appears in Section 24-2 of Dantzig's book, *Linear Programming and Extensions*.

simplex method of linear programming with a trivial modification." The suggestion alluded to by Künzi and Krelle can be found on the next-to-last page of the 23-page paper Markowitz (1956). There is no further mention of Markowitz or the CLA in Künzi and Krelle (1966).

In exercise 7-26 of his book *Nonlinear and Dynamic Programming* (Hadley (1964)), George Hadley introduced the portfolio selection problem, citing Markowitz (1959) rather than Markowitz (1952) as the source. Wolfe's simplex method (Wolfe (1959)), the Frank-Wolfe method (Frank and Wolfe (1956)), Beale's method (Beale (1959)), and Hildreth's method (Hildreth (1957)) are each discussed to some extent, but Markowitz's critical line method is not discussed at all.

The literature of quadratic programming has essentially only two monographs on the subject. J.C.G. Boot's book (Boot (1964)) devotes less than one page[2] of sketchy "conceptual" description to the Markowitz algorithm. Similarly, the book has little to say about Wolfe's "simplex algorithm for quadratic programming" (Wolfe (1959)). Boot does point out that Wolfe's paper deals with the case of a convex – as distinct from *strictly* convex – objective function, whereas in Markowitz's paper the latter possibility is no more than a footnote added in proof. See (Markowitz, 1956, p. 133). Instead, Boot presents what he calls Dantzig's Algorithm (the one found in Dantzig (1961)) at some length. On the one hand, as noted above, Dantzig assumes the objective function of his quadratic program is just a positive semidefinite quadratic form, rather than a more general quadratic function. Boot, on the other hand, gives the more general formulation, which makes a very important difference vis-à-vis the existence of an optimal solution and the possible unboundedness of the objective function. The other quadratic programming monograph van de Panne (1975), appeared 11 years after Boot's; except for one bibliographical entry, van de Panne makes no mention of Markowitz's work. By that time, interest in the CLA had all but vanished.

Years later, however, the algorithm was revisited in Markowitz's books Markowitz (1959) and Markowitz (1987). The latter was republished in 2000. This edition (Markowitz (2000)) is the same as Markowitz (1987) but contains a chapter by G.P. Todd presenting the CLA along with its implementation in a computer code.

With the possible exception of Perold (1984) on large-scale portfolio optimization, it would be an exaggeration to say that any of Markowitz's publications stimulated much research on the CLA per se within the *mathematical programming community.*

The remainder of this expository article consists of two sections. In Sect. 8.2 we speculate on the reasons for the remarkably limited interest in the CLA presented in Markowitz (1956) and Markowitz (1959) as an algorithm for quadratic programming. (This is not to suggest that it lacked influence in the world of portfolio selection.) It has been said in Markowitz (1959), Markowitz (1987), Markowitz (2002), and Perold (1984) that the CLA and Wolfe's simplex method for quadratic programming are equivalent. In Sect. 8.3 we describe an algorithm for parametric

---

[2] To be precise, it is page 181.

quadratic programming in the language of parametric linear complementarity theory. Interpreting the algorithms of Markowitz and Wolfe in terms of the one given here enables us to (re-)confirm that they are indeed equivalent. In so doing, we implicitly trace the impact of the CLA on the development of quadratic programming and linear complementarity.

8.2 Why of Limited Interest?

Markowitz's paper Markowitz (1956) can be said to have a number of qualities that discourage readership. The paper appeared in a journal, *Naval Research Logistics Quarterly*, which, for several years after its inception in 1954, presented articles in a typewriter-style font. One might think of blaming the physical appearance of the journal for turning interest away from Markowitz's paper, but that would be a mistake. Other papers such as Frank and Wolfe (1956) and Beale (1959) published in the same journal with the same printing process have gained greater recognition within the mathematical programming literature. A more credible explanation must be sought elsewhere; the paper's content and presentation style are the best places to start.

Compare the titles of Markowitz (1956) and Markowitz (1955). The former could be interpreted to refer to a *particular* quadratic function, whereas the latter seems broader in scope, though perhaps too much so. Furthermore, the use of "linear constraints" as distinct from "linear inequalities," leaves open the possibility that the paper is about the optimization of a quadratic function on a linear manifold, a topic that could be considered classical, even in the 1950s. Next comes the paper's abstract which, in its entirety, says

> *The author discusses a computational technique applicable to determination of the set of "efficient points" for quadratic programming problems.*

The term "quadratic programming" appears again in footnote 17, but nowhere else in the paper as far as we can see. The paper does, however, contain a few references to linear programming publications including Robert Dorfman's Ph.D. thesis/book Dorfman (1951) where the term "quadratic programming" was coined. Markowitz also mentions one of the two other quadratic programming works of the day: Houthakker (1953); however, he does not refer to the pioneering work of Barankin and Dorfman (1955). Markowitz could have cited the paper Frank and Wolfe (1956) which he, in fact, refereed (Markowitz, 2002, p. 155) and vice versa (Wolfe (2008)). But these are all petty quibbles.

Markowitz (1956) begins immediately with the problem of finding *efficient points* (E, V) as conceived in Markowitz (1952). The focus on efficient points (which is not especially characteristic of quadratic programming) is maintained through the first 11 of the paper's 12 sections. The 12th section, which appears on the penultimate page of the article, is called 'Minimizing a Quadratic." The opening paragraph of Sect. 12 is illuminating. It says

> One of the "by-products" of the calculation of efficient sets is the point at which V is a minimum, i.e., where $\lambda_E = 0$. The computing procedures described in Sections 6 through 10 are analogous to the simplex method of linear programming (as contrasted with the "gradient methods" that have been suggested[3] for both linear and non-linear programming). Both the procedure described in the preceding section—considered as a way of getting to min V—and the simplex method require a finite number of iterations, each iteration typically takes a "jump" to a new point which is superior to the old. *Each iteration makes use of the inverse of a matrix which is a "slight" modification of the matrix of the previous iteration. The success of the simplex method in linear programming suggests that it may be desirable to use a variant of the "critical line" in the quadratic case.*

(The italics are ours.)

Markowitz then translates the problem to one of minimizing a (presumably strictly convex) quadratic form $V = x'Cx$ subject to a system of general linear inequality constraints, which is to say subject to a system like

$$A^1x^1 + A^2x^2 = b^1,$$

$$A^3x^1 + A^4x^2 \geq b^2,$$

$$x^1 \geq 0, \quad x^2 \text{ free}.$$

Stating the constraints of the problem in such a general way has its computational merits, but it makes the exposition of the method unnecessarily difficult. The constraints could just as well have been expressed as an equivalent system of linear equations in nonnegative variables (as found in Wolfe (1959)). Doing so might have made a significant difference in how the paper was perceived.

The paper's notation also presents "stylistic challenges." For example, the intricate appearance of several equations (such as (Markowitz, 1956, eqs. (31), (32))) could have discouraged readers.

Perhaps most important of all the factors contributing to the paper's continued low profile is that other (contemporaneous) papers gave algorithms for *general* convex quadratic programming, even those for which a solution is not guaranteed. Although it is true that (in a footnote of Markowitz (1956)) the author claims the possibility of extending the procedures of his paper to instances of quadratic programming where the objective function includes a linear term and has a positive semidefinite Hessian matrix, he does not provide the details or address the issues that come up in such problems.

The monograph Markowitz (1959) improved the presentation considerably from the standpoint of typography, style, and underlying hypotheses. But having the title *Portfolio Selection*, the book could have failed to attract the quadratic programming audience we believe it deserved. Even if they discovered the material at the back of the book, many scholars interested in quadratic programming may not have wanted to grapple with so much material pertaining to the portfolio selection problem just to get to the heart of the matter. Nonetheless, Appendices A and B warrant serious

[3] Possibly an allusion to Frank and Wolfe (1956).

attention. Appendix A, "The Computation of Efficient Sets," gives a much more agreeable development of the critical line algorithm than Markowitz (1956) does. With some candor, Markowitz (Markowitz, 1959, p. 319) says of the *NRLQ* paper, "the exposition of the critical line method tends to become cluttered by tedious special cases. These 'special cases' lead to little, if any, change in computing procedure, but they must be covered to ensure that the method works in all circumstances." And, in the next paragraph he points out that "the treatment of a number of special cases in Markowitz (1956) did not depend on the restrictive assumptions about A and C. Those special cases will not be treated here." In Appendix B, "A Simplex Method for the Portfolio Selection Problem," Markowitz discusses the process of solving the portfolio selection problem by Wolfe's simplex method for quadratic programming, viewing it through the prism of the CLA. On the last page of this short appendix, Markowitz asserts the equivalence of the two algorithms and concludes by saying (Markowitz, 1959, p. 339) "The practical significance of the above result is that any of the linear programming codes for high speed, internally programmed computers can be conveniently converted to a quadratic programming code for solving, among other things, the portfolio selection problem."

8.3 On Algorithmic Equivalence

Is Markowitz's critical line algorithm equivalent to other algorithms for parametric quadratic programming, and in particular to Wolfe's simplex method? This question begs for clarification of the terminology.

What little the literature of mathematical programming has to say on the subject of algorithmic equivalence is rather informal. For relevant papers, see Best (1984), Cottle and Djang (1979), Goldfarb (1972), and Pang (1981). The doctoral dissertation Djang (1979) is an important source on this topic. As revealed in these publications, the essence of the concept is that two iterative algorithms are *equivalent* if – given the same inputs – they generate the same sequence of iterates. It should be added that two equivalent iterative algorithms should not only terminate with the same solution when it exists, but also provide a clear indication when it does not.

On a deeper level, another issue comes to mind. Must algorithmically equivalent algorithms do the same, or similar, amounts of work? This question is addressed in Djang (1979), but not in the other publications cited above; it is not addressed here either, but this is not to make light of its importance.

In comparing algorithms from the standpoint of equivalence, it is obviously essential to state clearly what sort of problem is being solved and what the steps of the algorithm are. These details become part of the overall assessment. For example, if one algorithm can be applied only to a problem having a strictly convex objective function or only to a problem having a bounded feasible region while the other algorithm applies to problems satisfying weaker conditions, it would generally be a mistake to call them equivalent in any but the more restrictive circumstances. As in the cases considered here, algorithms must be stated precisely. This should include whatever assumptions are made regarding degeneracy or the mechanisms they use for overcoming that condition.

Before proceeding to these formalities, we should point out that Markowitz has commented on the equivalence of the CLA and Wolfe's simplex method for quadratic programming. For example, he says in the aforementioned Appendix B (Markowitz, 1959, p. 339), "comparison of the amended simplex computation and the critical line procedure shows that, if they are started out together, they will continue together. ... The proof that the critical line method works, therefore, is a proof that the amended simplex method produces the desired results. It is this equivalence between the two procedures which implies that, when x_{js} and η_{js} are both outside the simplex basis, one will increase λ_E while the other will decrease it."

For Markowitz, the difference between the two algorithms is apparently a question of perspective. He puts the matter as follows (Markowitz, 1959, p. 222): "Wolfe was primarily concerned with minimizing a quadratic, and incidentally noted that the amended simplex algorithm would solve the portfolio selection problem; as compared with Markowitz, who was primarily concerned with the portfolio selection problem and incidentally noted that the CLA would minimize a quadratic."

The aim in this section is to establish the equivalence of Markowitz's CLA and Wolfe's Simplex Method for Quadratic Programming (SMQP) as applied to the *parametric quadratic programming (PQP) problem*

$$\begin{array}{ll} \text{minimize} & f(\lambda, x) = \lambda p'x + \frac{1}{2}x'Cx \\ \text{subject to} & Ax = b, \\ & x \geq 0. \end{array} \qquad (8.1)$$

A conventional (nonparametric) quadratic programming problem corresponds to the case where $\lambda = 1$. If $C = 0$, then (8.1) is a linear programming problem. If $\lambda=0$ or $p = 0$, it is the problem of minimizing a quadratic form (which is a special kind of quadratic programming problem).

Our approach to demonstrating the equivalence of the CLA and the SMQP is to relate each of them to a principal pivoting algorithm for such a parametric quadratic programming problem.

Assumptions and comments. We impose the following assumptions on the data.

(A1) The matrix C is symmetric and positive semidefinite.
A square matrix M can be called positive semidefinite as long as $x'Mx \geq 0$ for all x. The matrix M need not be symmetric for this to hold. Any square matrix M and its *symmetric part*, $\frac{1}{2}(M + M')$, satisfy the identity

$$x'Mx = \tfrac{1}{2}x'(M + M')x \quad \text{for all } x.$$

Later in this development, we shall encounter *nonsymmetric* positive semidefinite matrices.

(A2) $A \in R^{m \times n}$.
This assumption simply states the dimensions of the matrix A, and determines those of C, b, p, and x.

(A3) $\text{rank}(A) = m$.
This means that $m \leq n$ and the rows of A are linearly independent. This mild assumption is easily checked by methods of linear algebra. If the rank of A is $r < m$, then $m - r$ equations can be discarded from the system $Ax = b$.

(A4) The feasible region $\mathcal{X} := \{x : Ax = b,\ x \geq 0\}$ is nonempty.
Feasibility can be checked by methods of linear programming. Doing this is actually a part of the CLA and SMQP. If (A4) is not satisfied, there cannot be a solution to the problem.

(A5) The feasible region $\mathcal{X}$ is bounded.
The boundedness property, while not altogether standard in quadratic programming lore, is typical of portfolio selection problems because of the budget constraint and the nonnegativity of the variables (no short positions). As a practical computational matter, very large upper bounds can be imposed on individual variables, thereby insuring that (A5) is satisfied. The combination of (A4) and (A5) assures the existence of an optimal solution to the problem for every $\lambda \geq 0$.

(A6) Nondegeneracy holds when needed.
This assumption is imposed for ease of exposition. In point of fact, there are several methods (such as perturbation, lexicographic ordering, least-index selection rules) for dealing with degeneracy; describing them in detail is beyond the scope of this article. See Dantzig et al. (1955), Graves (1967), and Chang and Cottle (1980) among others.

Before turning to the advertised business, we bring up an elementary observation, which – while not essential – is interesting because when its hypothesis holds, there is no need for assumption (A5) to assert the existence of a solution to (8.1).

Proposition 8.1. *If the linear program*

$$\begin{aligned} &\textit{minimize} && p'x \\ &\textit{subject to} && Ax = b, \\ &&& x \geq 0 \end{aligned} \tag{8.2}$$

has an optimal solution, then, for every $\lambda \geq 0$*, so does the quadratic program*

$$\begin{aligned} &\textit{minimize} && \lambda p'x + \tfrac{1}{2}x'Cx \\ &\textit{subject to} && Ax = b, \\ &&& x \geq 0. \end{aligned} \tag{8.3}$$

Proof. If $\bar{x}$ is an optimal solution of (8.2), then, for every $\lambda \geq 0$, the vector $\bar{x}$ minimizes $\lambda p'x$ subject to $x \in \mathcal{X}$. For fixed $\lambda \geq 0$ function $f(\lambda, x)$ is bounded below on the polyhedral feasible region $\mathcal{X}$ of the QP because of the existence of an optimal solution for the LP and the fact that $x'Cx \geq 0$ for all x. The so-called Frank-Wolfe theorem (Frank and Wolfe, 1956, Appendix (i)) then guarantees the existence of a minimum for the quadratic program. $\square$

It is worth noting that in the absence of assumption (A5), the existence of an optimal solution to (8.3) does *not* imply the existence of an optimal solution to (8.2).

8.3.1 First-Order Optimality (KKT) Conditions

In dealing with any optimization problem, it is essential to have a clear understanding of how to recognize an optimal solution. In nonlinear programming where differentiability and certain "regularity conditions" are satisfied, the Karush–Kuhn–Tucker (KKT) theorem (Karush (1939), Kuhn and Tucker (1951)) is instrumental in providing such an understanding. Nearly every paper ever written on quadratic programming uses the KKT conditions in one way or another. The early publications Barankin and Dorfman (1955), Barankin and Dorfman (1956), Barankin and Dorfman (1958), Dennis (1959), Frank and Wolfe (1956), Hildreth (1954), Hildreth (1957), Markowitz (1956), Markowitz (1959), and Wolfe (1959) all exemplify the tradition. We invoke the usual first-order KKT conditions of optimality by forming the *Lagrangian function* associated with the PQP (8.1) namely

$$\mathcal{L}(\lambda, x, u, v) = \lambda p'x + \tfrac{1}{2}x'Cx + u'(Ax - b) - v'x.$$

The necessary conditions of optimality are then expressed as

$$\nabla_x \mathcal{L}(\lambda, x, u, v) = \lambda p + Cx + A'u - v = 0 \tag{8.4}$$

$$\nabla_u \mathcal{L}(\lambda, x, u, v) = Ax - b = 0 \tag{8.5}$$

$$x \geq 0, \quad v \geq 0, \quad u \text{ free} \tag{8.6}$$

$$x'v = 0 \tag{8.7}$$

Assumption (A1) makes the function $f(\lambda, x)$ *convex* in x for every fixed $\lambda \geq 0$; for this reason these KKT conditions are sufficient as well as necessary for optimality.

The KKT theorem dates from an era when the simplex method for linear programming was a hot new topic in mathematics and economics. The observation in Barankin and Dorfman (1955) that the stationarity conditions (8.4) and (8.5) are linear equations inspired efforts to solve the entire system (8.4)–(8.7) by methods akin to the simplex method of linear programming.

To enforce the feasibility of a nonnegative vector x, we find it convenient to define

$$y := b - Ax \tag{8.8}$$

and then require y to be zero. This will be accomplished by making y nonbasic in a system of equations.

With the aid of (8.8), the KKT conditions for our PQP can be expressed (in the so-called "dictionary form") as

$$v = \lambda p + Cx + A'u, \tag{8.9}$$

$$y = b - Ax, \tag{8.10}$$

$$x \geq 0, \quad v \geq 0, \quad y = 0, \quad u \text{ free}, \tag{8.11}$$

$$x'v = 0. \tag{8.12}$$

We seek a solution to this system for every $\lambda \geq 0$.

The system given by (8.9), (8.10), (8.11), and (8.12) is an instance of a *parametric mixed linear complementarity problem* (PMLCP). With $\mathrm{N} = \{1, \ldots, n\}$ and $\mathrm{K} = \{n+1, \ldots, n+m\}$, this PMLCP can be written in the form

$$w = q + \lambda d + Mz, \tag{8.13}$$

$$w_{\mathrm{N}} \geq 0, \quad z_{\mathrm{N}} \geq 0, \quad w_{\mathrm{K}} = 0, \quad z_{\mathrm{K}} \text{ free}, \tag{8.14}$$

$$z'_{\mathrm{N}} w_{\mathrm{N}} = 0, \tag{8.15}$$

where

$$w = \begin{bmatrix} v \\ y \end{bmatrix}, \quad z = \begin{bmatrix} x \\ u \end{bmatrix}, \quad q = \begin{bmatrix} 0 \\ b \end{bmatrix}, \quad d = \begin{bmatrix} p \\ 0 \end{bmatrix}, \quad \text{and} \quad M = \begin{bmatrix} C & A' \\ -A & 0 \end{bmatrix}. \tag{8.16}$$

The matrix M in (8.16) possesses a special property known as *bisymmetry*. This is best seen in (8.16). The block structure of M is a combination of ordinary symmetry and skew-symmetry. Indeed,

$$M = \begin{bmatrix} C & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & A' \\ -A & 0 \end{bmatrix}.$$

The first summand is symmetric, and the second is skew-symmetric.

Note that $q \in R^L$, $d \in R^L$, and $M \in R^{L \times L}$, where $L = n + m$. In the system (8.13), the variables $w_1, \ldots, w_L$ play the role of *basic variables*, whereas the variables $z_1, \ldots, z_L$ are *nonbasic variables*. As we shall see, the role played by a variable can change through the process of re-representing the system. But it is always true that the basic variables are *dependent* on the *independent* nonbasic variables, whatever those may be.

When the index set K is vacuous, the system above is a *standard parametric linear complementarity problem.* We drop the word "parametric" when $d = 0$ in either the mixed or standard PLCP. In either case (mixed or standard) when λ is constant, we have a *nonparametric* MLCP or LCP.

Note that for any positive semidefinite matrix C, the matrix M as defined in (8.16) above is positive semidefinite; moreover, M is nonsymmetric if and only if $A \neq 0$.

The following theorem is well known.

Theorem 8.1. *A standard parametric linear complementarity problem (PLCP) with data q and M, where M is positive semidefinite, has a solution if (and only if) there exist $w \geq 0$ and $z \geq 0$ such that $w = q + Mz$.*

Proof. See (Cottle et al., 1992, p. 139). □

It is a simple matter to show that this theorem is also valid for a mixed nonparametric linear complementarity problem. So for fixed λ, the existence of a solution to the necessary and sufficient conditions of optimality in our quadratic programming problem rests on the *feasibility* of the system of linear inequalities, (8.9)–(8.11).

For either the mixed or standard problem, *the range of λ values for which the aforementioned system of linear inequalities is feasible can be determined by solving a pair of linear programming problems in which λ is first maximized and then minimized subject to the linear inequalities.* In our case, these linear programming problems are to find

$$\lambda_{\min} = \inf\{\lambda \geq 0 : (8.9),\ (8.10)\ \text{and}\ (8.11)\ \text{are feasible}\},$$

$$\lambda_{\max} = \sup\{\lambda \geq 0 : (8.9),\ (8.10)\ \text{and}\ (8.11)\ \text{are feasible}\}.$$

In the case at hand, once we find a basic feasible solution for the constraints given in (8.1), we can declare that $\lambda_{\min} = 0$. And as for $\lambda_{\max}$, we need not solve the linear programming problem that reveals its value. The principal pivoting algorithm described below will automatically do so.

Tableau form of the (KKT) system. The conditions (8.9)–(8.11) can be represented in tableau form as

$$\begin{array}{c|c|c|cc|c} & 1 & \lambda & x & u & \\ \hline v & 0 & p & C & A' & \oplus \\ y & b & 0 & -A & 0 & 0 \\ \hline & 1 & \oplus & \oplus & \star & \end{array} \tag{8.17}$$

The symbols along the right edge and bottom of the tableau serve as reminders of the desired properties of those along the left and top of the tableau. The symbol $\oplus$ stands for "nonnegative." The 0 on the right edge means that the corresponding vector (in this case y) is required to be 0. The symbol $\star$ means that there is no sign or value restriction on the corresponding vector (in this case u).

Notational matters. Unless stated otherwise, we regard all vectors as columns, that is, 1-column matrices. The notational convention $x = (x_1, \ldots, x_n)$ is simply meant to save vertical space and avoid unnecessary transposes.

We have already defined $\mathrm{N} = \{1, \dots, n\}$. This denotes the full set of indices used to label the columns of A. Let J denote the index set $\{j_1, \dots, j_m\}$ corresponding to the columns of a *basis* in A. We can represent the basis in A by the symbol $A_{\bullet\mathrm{J}}$. (The dot after A signifies "all rows.") Relative to the system $Ax = b$, $x \geq 0$, a *feasible basis* $A_{\bullet\mathrm{J}}$ is one for which $A_{\bullet\mathrm{J}}^{-1}b \geq 0$. Let x_{J} denote the subvector of x corresponding to the index set J. Thus, we have $x_{\mathrm{J}} = (x_{j_1}, \dots, x_{j_m})$. Next, let H denote the complement of J with respect to N. This means that H and J form a partition of N, that is, $\mathrm{N} = \mathrm{H} \cup \mathrm{J}$ and $\mathrm{H} \cap \mathrm{J} = \emptyset$.

In systems such as (8.9) and (8.10) or their tableau form (8.17), the coefficients of an independent (nonbasic) variable can be viewed as partial derivatives of corresponding dependent (basic) variables. Thus, in (8.17) for example, the coefficient of x_j in the row of y_i can be expressed as $\partial y_i / \partial x_j$. We employ this representation repeatedly.

For ease of notation, we assume that $A = [A_{\bullet\mathrm{H}}\ A_{\bullet\mathrm{J}}]$. The same partitioning can be extended to the other data of the problem. Thus we have

$$x = (x_{\mathrm{H}}, x_{\mathrm{J}}), \quad p = (p_{\mathrm{H}}, p_{\mathrm{J}}), \quad v = (v_{\mathrm{H}}, v_{\mathrm{J}})$$

and

$$C = \begin{bmatrix} C_{\mathrm{HH}} & C_{\mathrm{HJ}} \\ C_{\mathrm{JH}} & C_{\mathrm{JJ}} \end{bmatrix}.$$

The operations of transposition and inversion applied to submatrices are applied after the extraction. Thus, $A'_{\bullet\mathrm{H}}$ means $(A_{\bullet\mathrm{H}})'$, and $A_{\bullet\mathrm{J}}^{-1}$ means $(A_{\bullet\mathrm{J}})^{-1}$. This convention avoids some unnecessary parentheses. Even so, parentheses are advisable in cases like $(A'_{\bullet\mathrm{J}})^{-1}$.

Using this notational scheme, we can write the tableau (8.17) as

$$\begin{array}{c|c|c|ccc|c} & 1 & \lambda & x_{\mathrm{H}} & x_{\mathrm{J}} & u & \\ \hline v_{\mathrm{H}} & 0 & p_{\mathrm{H}} & C_{\mathrm{HH}} & C_{\mathrm{HJ}} & A'_{\bullet\mathrm{H}} & \oplus \\ v_{\mathrm{J}} & 0 & p_{\mathrm{J}} & C_{\mathrm{JH}} & C_{\mathrm{JJ}} & A'_{\bullet\mathrm{J}} & \oplus \\ y & b & 0 & -A_{\bullet\mathrm{H}} & -A_{\bullet\mathrm{J}} & 0 & 0 \\ \hline & 1 & \oplus & \oplus & \oplus & \star & \end{array} \tag{8.18}$$

Lemma 8.1. *If $A_{\bullet\mathrm{J}}$ is a basis in A, the matrix*

$$P = \begin{bmatrix} C_{\mathrm{JJ}} & A'_{\bullet\mathrm{J}} \\ -A_{\bullet\mathrm{J}} & 0 \end{bmatrix}$$

is nonsingular, and

$$P^{-1} = \begin{bmatrix} 0 & -A_{\bullet\mathrm{J}}^{-1} \\ (A'_{\bullet\mathrm{J}})^{-1} & (A'_{\bullet\mathrm{J}})^{-1} C_{\mathrm{JJ}} A_{\bullet\mathrm{J}}^{-1} \end{bmatrix}.$$

Proof. Multiply the two given matrices. □

We remark in passing that both P and P^{-1} are bisymmetric.

It follows that the matrix P can be used to carry out a *block pivot*. This operation has the effect of exchanging the corresponding subsets of basic variables and nonbasic variables as well as transforming the data in the tableau. In particular, the nonsingularity of the matrix P above allows us to exchange the nonbasic vector (x_{J}, u) with the basic vector (v_{J}, y). To make this process a little easier to comprehend, we define the index set $\mathrm{K} = \{n+1, \dots, n+m\}$ and let $v_{\mathrm{K}} = y$. Thus $v_{n+i} = y_i$ for $i = 1, \dots, m$. Likewise, let $x_{\mathrm{K}} = u$, so that $x_{n+i} = u_i$ for $i = 1, \dots, m$.

With obvious identifications, the tableau (8.18) can then be written

$$\begin{array}{c|c|c|ccc|c} & 1 & \lambda & x_{\mathrm{H}} & x_{\mathrm{J}} & x_{\mathrm{K}} & \\ \hline v_{\mathrm{H}} & q_{\mathrm{H}} & d_{\mathrm{H}} & M_{\mathrm{HH}} & M_{\mathrm{HJ}} & M_{\mathrm{HK}} & \oplus \\ v_{\mathrm{J}} & q_{\mathrm{J}} & d_{\mathrm{J}} & M_{\mathrm{JH}} & M_{\mathrm{JJ}} & M_{\mathrm{JK}} & \oplus \\ v_{\mathrm{K}} & q_{\mathrm{K}} & d_{\mathrm{K}} & M_{\mathrm{KH}} & M_{\mathrm{KJ}} & M_{\mathrm{KK}} & 0 \\ \hline & 1 & \oplus & \oplus & \oplus & \star & \end{array} \tag{8.19}$$

Pivoting in the tableau on the block P can be viewed as the process whereby the equation for v_{K} is solved for x_{J}, the equation for v_{J} is solved for x_{K}, and then the corresponding rows and columns are permuted so as to preserve the numerical order of the subscripts. The resulting tableau will look as follows

$$\begin{array}{c|c|c|ccc|c} & 1 & \lambda & x_{\mathrm{H}} & v_{\mathrm{J}} & v_{\mathrm{K}} & \\ \hline v_{\mathrm{H}} & \bar{q}_{\mathrm{H}} & \bar{d}_{\mathrm{H}} & \bar{M}_{\mathrm{HH}} & \bar{M}_{\mathrm{HJ}} & \bar{M}_{\mathrm{HK}} & \oplus \\ x_{\mathrm{J}} & \bar{q}_{\mathrm{J}} & \bar{d}_{\mathrm{J}} & \bar{M}_{\mathrm{JH}} & \bar{M}_{\mathrm{JJ}} & \bar{M}_{\mathrm{JK}} & \oplus \\ x_{\mathrm{K}} & \bar{q}_{\mathrm{K}} & \bar{d}_{\mathrm{K}} & \bar{M}_{\mathrm{KH}} & \bar{M}_{\mathrm{KJ}} & \bar{M}_{\mathrm{KK}} & \star \\ \hline & 1 & \oplus & \oplus & \oplus & 0 & \end{array} \tag{8.20}$$

Making the precise formulas for the 6 vector- and 9 matrix-entries fit in the new tableau presents a typographical challenge (if readability is a criterion), so instead we simply list them.

$$\bar{q}_{\mathrm{H}} = -A'_{\bullet\mathrm{H}}(A'_{\bullet\mathrm{J}})^{-1}C_{\mathrm{JJ}}A_{\bullet\mathrm{J}}^{-1}b + C_{\mathrm{HJ}}A_{\bullet\mathrm{J}}^{-1}b \tag{8.21}$$

$$\bar{q}_{\mathrm{J}} = A_{\bullet\mathrm{J}}^{-1}b \tag{8.22}$$

$$\bar{q}_{\mathrm{K}} = -(A'_{\bullet\mathrm{J}})^{-1}C_{\mathrm{JJ}}A_{\bullet\mathrm{J}}^{-1}b \tag{8.23}$$

$$\bar{d}_{\mathrm{H}} = d_{\mathrm{H}} - A'_{\bullet\mathrm{H}}(A'_{\bullet\mathrm{J}})^{-1}d_{\mathrm{J}} \tag{8.24}$$

$$\bar{d}_{\mathrm{J}} = 0 \tag{8.25}$$

$$\bar{d}_{\mathrm{K}} = -(A'_{\bullet\mathrm{J}})^{-1}d_{\mathrm{J}} \tag{8.26}$$

$$\bar{M}_{\mathrm{HH}} = C_{\mathrm{HH}} - C_{\mathrm{HJ}}A_{\bullet\mathrm{J}}^{-1}A_{\bullet\mathrm{H}} - A'_{\bullet\mathrm{H}}(A'_{\bullet\mathrm{J}})^{-1}C_{\mathrm{JH}} + A'_{\bullet\mathrm{H}}(A'_{\bullet\mathrm{J}})^{-1}C_{\mathrm{JJ}}A_{\bullet\mathrm{J}}^{-1}A_{\bullet\mathrm{H}} \tag{8.27}$$

$$\bar{M}_{\mathrm{HJ}} = A'_{\bullet\mathrm{H}}(A'_{\bullet\mathrm{J}})^{-1} \tag{8.28}$$

$$\bar{M}_{\mathrm{HK}} = -C_{\mathrm{HJ}} A_{\bullet\mathrm{J}}^{-1} + A'_{\mathrm{H}} (A'_{\bullet\mathrm{J}})^{-1} C_{\mathrm{JJ}} A_{\bullet\mathrm{J}}^{-1} \tag{8.29}$$

$$\bar{M}_{\mathrm{JH}} = -A_{\bullet\mathrm{J}}^{-1} A_{\bullet\mathrm{H}} \tag{8.30}$$

$$\bar{M}_{\mathrm{JJ}} = 0 \tag{8.31}$$

$$\bar{M}_{\mathrm{JK}} = -A_{\bullet\mathrm{J}}^{-1} \tag{8.32}$$

$$\bar{M}_{\mathrm{KH}} = -(A'_{\bullet\mathrm{J}})^{-1} C_{\mathrm{JH}} + (A'_{\bullet\mathrm{J}}) C_{\mathrm{JJ}} A_{\bullet\mathrm{J}}^{-1} A_{\bullet\mathrm{H}} \tag{8.33}$$

$$\bar{M}_{\mathrm{KJ}} = (A'_{\bullet\mathrm{J}})^{-1} \tag{8.34}$$

$$\bar{M}_{\mathrm{KK}} = (A'_{\bullet\mathrm{J}})^{-1} C_{\mathrm{JJ}} A_{\bullet\mathrm{J}}^{-1} \tag{8.35}$$

Although it may not be apparent from the transformed data given above, it is nonetheless a fact that the matrix $\bar{M}$ in the new tableau is positive semidefinite. The new tableau is obtained from its predecessor by what is called a *principal pivot transformation.* See Tucker (1963) and (Cottle et al., 1992, pp. 71–72) for the formulas through which these entries are defined.

Theorem 8.2. *The classes of positive definite and positive semidefinite matrices are invariant under principal pivoting.*

Proof. See (Cottle et al., 1992, **4.1.5** and **4.12.1**). □

Principal pivoting also preserves bisymmetry as shown in Keller (1973). These two invariance theorems on principal pivoting constitute a key part of the theoretical foundation for the principal pivoting approach to solving the parametric quadratic programming as formulated above. Another is the following simple result.

Proposition 8.2. *Let C be an $n \times n$ symmetric positive semidefinite matrix. If for any index i the diagonal element $c_{ii} = 0$, then $c_{ij} = c_{ji} = 0$ for all $j = 1, \ldots, n$.*

Proof. See (Cottle, 1968, Theorem 3). □

A further observation can be made about the tableau (8.20). Since the (now basic) vector x_{K} is unrestricted in sign, and (now nonbasic) vector v_{K} must be 0, the rows and columns associated with each can be ignored or even dropped from the tableau.[4] Once that is done, what remains is a standard PLCP as represented by the tableau

$$\begin{array}{c|c|c|cc|c} & 1 & \lambda & x_{\mathrm{H}} & v_{\mathrm{J}} & \\ \hline v_{\mathrm{H}} & \bar{q}_{\mathrm{H}} & \bar{d}_{\mathrm{H}} & \bar{M}_{\mathrm{HH}} & \bar{M}_{\mathrm{HJ}} & \oplus \\ x_{\mathrm{J}} & \bar{q}_{\mathrm{J}} & \bar{d}_{\mathrm{J}} & \bar{M}_{\mathrm{JH}} & \bar{M}_{\mathrm{JJ}} & \oplus \\ \hline & 1 & \oplus & \oplus & \oplus & \end{array} \tag{8.36}$$

[4] The first part of this observation was made in Wolfe (1959).

The matrix of this tableau is also bisymmetric and positive semidefinite. A solution to the reduced problem requires all n pairs of the remaining variables x_j and v_j to be nonnegative and complementary, i.e., satisfy

$$x_j \geq 0, \quad v_j \geq 0, \quad \text{and} \quad x_j v_j = 0 \quad \text{for all } j = 1, \dots, n.$$

Once the problem is solved, the Lagrange multiplier vector $u = u(\lambda)$ can be recovered from the formula for v_{K} implicit in (8.19).

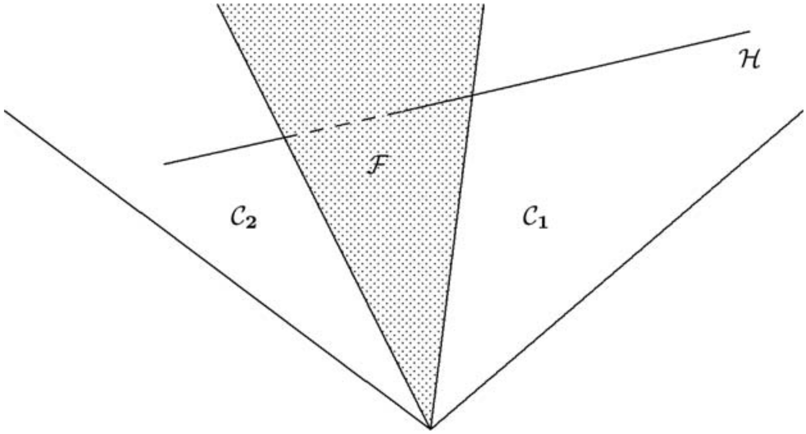
8.3.2 Solving the PQP (36) and Its Geometric Interpretation

The methodology presented here for solving the PQP (8.1) has two aspects. After obtaining a basic feasible solution of the LP in (8.2), and obtaining the PLCP corresponding to the tableau (8.20), it starts with $\lambda = 0$ and a complementary, but possibly infeasible, solution (x, v) to the constraints. The infeasibility (if any) is concentrated in the subvector v_{H} which may have one or more negative components. The goal is to find a partition of N into sets J and H such that the associated complementary basic solution is feasible. Considering where the procedure starts, it is a matter of making the subvector v_{H} become nonnegative. Once this is accomplished, the parametric part (the increase of λ) begins; and the task is to retain the nonnegativity of the basic variables by finding a new complementary feasible basis if and when λ reaches a new breakpoint.

There is a useful geometric interpretation of the parametric aspect of this algorithm. When the problem corresponding to $\lambda = 0$ is solved, the vector $(\bar{q}_{\mathrm{H}}, \bar{q}_{\mathrm{J}})$ will have been expressed as an element of a *complementary cone* relative to the matrix

$$\widehat{M} = \begin{bmatrix} \bar{M}_{\mathrm{HH}} & \bar{M}_{\mathrm{HJ}} \\ \bar{M}_{\mathrm{JH}} & \bar{M}_{\mathrm{JJ}} \end{bmatrix}$$

drawn from the tableau shown in (8.36). A complementary cone $\mathcal{C}$ is the set of all vectors that can be expressed as nonnegative linear combinations of the columns of a *complementary submatrix*. Relative to $\widehat{M}$, a complementary submatrix B is one such that for every $j = 1, \dots, n$, the jth column of B is either $I_{\bullet j}$ or $-\widehat{M}_{\bullet j}$. The union of the finitely many complementary cones relative to $\widehat{M}$ is called its *complementary range* and is denoted $K(\widehat{M})$. The complementary range of $\widehat{M}$ is the union of finitely many closed convex cones (each of which is finitely generated) and hence is a closed cone. It is known that when $\widehat{M}$ is positive semidefinite, its complementary range is convex. The convexity of $K(\widehat{M})$ implies that the PLCP corresponding to (8.36) has a unique solution for all λ in the interval $[\lambda_{\min}, \lambda_{\max}]$. By (A6) and the positive semidefiniteness of $\widehat{M}$, the solution is unique. (See (Lemke, 1965, p. 685) and Cottle and Stone (1983).) As the parameter λ increases, the *half-line* $\mathcal{H} = \{r(\lambda) = (\bar{q}_{\mathrm{H}}, \bar{q}_{\mathrm{J}}) + \lambda(\bar{d}_{\mathrm{H}}, \bar{d}_{\mathrm{J}}) : \lambda \geq 0\}$ traverses a particular complementary cone. The corresponding tabular form makes it easy to read off the values of the basic variables. By the nondegeneracy assumption, the basic variables will be positive except when the point on the half-line corresponding to a particular value of λ lies on the boundary of the current complementary cone. When such a thing occurs, *only one* basic variable will equal zero. A further increase of λ would require that basic variable to become *negative*, meaning that the corresponding point on the half-line no longer belongs to the complementary cone with which the representation in the current tableau is associated. This signals the need to change the basis. Thus, if, for $h \in \mathrm{H}$, v_h decreases to zero, its place it taken by x_h, in which case h is removed from H and adjoined to J. The corresponding pivot operation is called an IN-PIVOT. If $j \in \mathrm{J}$, it is x_j that first decreases to zero, then v_j becomes a basic variable and j is taken out of H and put into H. In this case, the corresponding pivot operation is called an OUT-PIVOT. For further discussion of complementary cones and submatrices, see Cottle et al. (1992).



Half-line $\mathcal{H}$ passing through adjacent complementary cones.
Shadded cone $\mathcal{F}$ represents the facet common to the adjacent cones $\mathcal{C}_1$ and $\mathcal{C}_2$

Actually, in the problem at hand, $\lambda_{\min} = 0$, because the objective function of the PQP is bounded below on its nonempty feasible region. On the other hand, $\lambda_{\max}$ can be either finite or $+\infty$. The latter alternative would hold whenever the linear program (8.2) has an optimal solution, which it must if (A4) and (A5) hold.

In light of our earlier comments about eliminating the vectors x_{K} and v_{K}, the first task is to find a solution to the standard linear complementarity problem

$$v_{\mathrm{H}} = \bar{q}_{\mathrm{H}} + \bar{M}_{\mathrm{HH}} x_{\mathrm{H}} + \bar{M}_{\mathrm{HJ}} v_{\mathrm{J}}, \tag{8.37}$$

$$x_{\mathrm{J}} = \bar{q}_{\mathrm{J}} + \bar{M}_{\mathrm{JH}} x_{\mathrm{H}} + \bar{M}_{\mathrm{JJ}} v_{\mathrm{J}}, \tag{8.38}$$

$$x_{\mathrm{H}} \geq 0, \;\; x_{\mathrm{J}} \geq 0, \;\; v_{\mathrm{H}} \geq 0, \;\; v_{\mathrm{J}} \geq 0, \tag{8.39}$$

$$x'_{\mathrm{H}} v_{\mathrm{H}} = x'_{\mathrm{J}} v_{\mathrm{J}} = 0. \tag{8.40}$$

Note. This LCP corresponds to solving the parametric problem with $\lambda = \lambda_{\min} = 0$. Starting the procedure at $\lambda = \lambda_{\max}$ can also be done, but then the vector $(\bar{q}_{\mathrm{H}}, \bar{q}_{\mathrm{J}})$

would be replaced by $(\bar{q}_{\mathrm{H}}, \bar{q}_{\mathrm{J}}) + \lambda_{\max}(\bar{d}_{\mathrm{H}}, \bar{d}_{\mathrm{J}})$. If $\lambda_{\max} = +\infty$, a sufficiently large finite value can be used. This can be reformulated by letting $\theta = \lambda_{\max} - \lambda \geq 0$. Then $\lambda = \lambda_{\max} - \theta$. Substitution into $(\bar{q}_{\mathrm{H}}, \bar{q}_{\mathrm{J}}) + \lambda(\bar{d}_{\mathrm{H}}, \bar{d}_{\mathrm{J}})$ gives

$$(\bar{q}_{\mathrm{H}}, \bar{q}_{\mathrm{J}}) + \lambda_{\max}(\bar{d}_{\mathrm{H}}, \bar{d}_{\mathrm{J}}) - \theta(\bar{d}_{\mathrm{H}}, \bar{d}_{\mathrm{J}}) = (\bar{q}_{\mathrm{H}} + \lambda_{\max}\bar{d}_{\mathrm{H}}, \bar{q}_{\mathrm{J}} + \lambda_{\max}\bar{d}_{\mathrm{J}}) + \theta(-\bar{d}_{\mathrm{H}}, -\bar{d}_{\mathrm{J}}).$$

This has essentially the same form as where we start from $\lambda_{\min} = 0$ but with a different base vector $\bar{q}$ and a different direction vector $\bar{d}$.

It is worth noting that in the system of (8.37) and (8.38), we have a basic (though not necessarily feasible) solution. The variables in v_{H} and x_{J} are basic; their complements, the variables in x_{H} and v_{J}, are nonbasic. Moreover, all three types of pivot operations described below preserve this property as well as the property that $x_i v_i = 0$ for all $i \in \mathrm{N}$. The object of the algorithm is to achieve a complementary basic solution that is *nonnegative*, hence feasible. This can be done only if the LCP has a solution. If our hypotheses do not guarantee the existence of a solution to the QP and hence to the LCP, the statement of the algorithm needs to provide for this contingency, and it does so.

Principal pivoting algorithm for solving the LCP (36) with $\boldsymbol{\lambda = \lambda_{\min} = 0}$. The nonbasic vectors x_{H} and v_{J} are set at zero. From the initialization step, we know that $\bar{q}_{\mathrm{J}} \geq 0$, hence $x_{\mathrm{J}} \geq 0$. If $v_{\mathrm{H}} = \bar{q}_{\mathrm{H}} \geq 0$, a solution of the LCP is at hand.

Suppose $\bar{q}_{\mathrm{H}}$ contains a negative component, say $\bar{q}_i$.[5] As noted earlier, the matrix

$$\begin{bmatrix} \bar{M}_{\mathrm{HH}} & \bar{M}_{\mathrm{HJ}} \\ \bar{M}_{\mathrm{JH}} & \bar{M}_{\mathrm{JJ}} \end{bmatrix}$$

is positive semidefinite. From (8.27), (8.28), (8.30), and (8.31) this matrix is easily seen to be bisymmetric:

$$\bar{M}_{\mathrm{HH}} = \bar{M}'_{\mathrm{HH}}, \quad \bar{M}_{\mathrm{JJ}} = \bar{M}'_{\mathrm{JJ}}, \quad \text{and} \quad -\bar{M}_{\mathrm{HJ}} = \bar{M}'_{\mathrm{JH}}.$$

In this case, noting that $\bar{m}_{ii} = \partial v_i / \partial x_i \geq 0$, we increase the nonbasic variable x_i. There are several possibilities:

1. $\bar{m}_{ii} > 0$ and the increase of x_i does not cause any basic variable x_j to decrease to zero before v_i increases to zero. In this case, we pivot on $\bar{m}_{ii}$. This is called an In-Pivot. It transfers the index i from H into the set J.
2. $\bar{m}_{ii} \geq 0$ and the increase of x_i causes some basic variable x_j to decrease to zero before v_i increases to zero. In this instance, there are two possibilities.

   (a) $\bar{m}_{jj} > 0$. In this case, we can pivot on $\bar{m}_{jj}$ thereby making v_j basic in place of x_j. When this happens, we transfer j from J to H. This pivot operation is called an Out-Pivot. After this pivot occurs we attempt to increase the basic variable v_i by resuming the increase of x_i.

[5] There could be several indices i such that $\bar{q}_i < 0$. The matter of which one to select is analogous to choosing the "incoming" column in the simplex method for linear programming.

(b) $\bar{m}_{jj} = 0$. We know that $\bar{m}_{ji} = \partial x_j / \partial x_i < 0$. From the bisymmetry property, we have $\bar{m}_{ij} = \partial v_i / \partial v_j > 0$. This implies that

$$\begin{bmatrix} \bar{m}_{ii} & \bar{m}_{ij} \\ \bar{m}_{ji} & \bar{m}_{jj} \end{bmatrix}$$

is nonsingular. It can serve as a block pivot for what is called an Exchange-Pivot. Exchange-Pivots change the composition of the index sets H and J, but not their cardinality. Note that after such a pivot, the distinguished variable v_i becomes nonbasic at a negative value. In this case, it is to be increased; but while it is nonbasic, it is not to be increased beyond zero. Once it reaches zero, the condition $x'v = 0$ will have been restored.

3. $\bar{m}_{ii} = 0$ and $\bar{m}_{ji} \geq 0$ for all $j \in \mathrm{J}$. If this occurs, the ith column of $\widehat{M}$ is nonnegative, and the ith row of $\widehat{M}$ is nonpositive. The assumption that $\bar{q}_i < 0$ implies that the ith constraint

$$v_i = \bar{q}_i + \sum_{h \in \mathrm{H}} \bar{m}_{ih} x_h + \sum_{j \in \mathrm{J}} \bar{m}_{ij} v_j = \bar{q}_i + \sum_{j \in \mathrm{J}} \bar{m}_{ij} v_j \geq 0$$

cannot be satisfied with $x_{\mathrm{H}} \geq 0$ and $v_{\mathrm{J}} \geq 0$. In the present circumstances, this outcome does not occur since the KKT conditions must have a solution.

It is known that, in the nondegenerate case, In-Pivots, Out-Pivots, and Exchange-Pivots all decrease the objective function value. Since the choice of a basis uniquely determines the tableau (up to rearrangements of the rows and columns), the pivoting method outlined above terminates in finitely many steps with vector x that minimizes $f(0, x)$ subject to $x \in \mathcal{X}$.

We are now in a position to proceed to the parametric part of the algorithm.

Increasing $\boldsymbol{\lambda}$*: the parametric part.* Assume now that the PLCP has been solved for $\lambda = \lambda_{\min} = 0$. We will then have a principal pivot transform of the system given by (8.37) and (8.38). Because we are going to be increasing λ from its current value of zero, we need to insert the corresponding column, which would be the pivotal transform of the subvectors of $\bar{d}$ given in (8.24) and (8.25). These could have been obtained by leaving the appropriate column in the model for $\lambda = 0$ and then transforming it along with the other data.

Suppose the procedure just described yields the system

$$v_{\mathrm{H}} = \hat{q}_{\mathrm{H}} + \lambda \hat{d}_{\mathrm{H}} + \hat{M}_{\mathrm{HH}} x_{\mathrm{H}} + \hat{M}_{\mathrm{HJ}} v_{\mathrm{J}}, \tag{8.41}$$

$$x_{\mathrm{J}} = \hat{q}_{\mathrm{J}} + \lambda \hat{d}_{\mathrm{J}} + \hat{M}_{\mathrm{JH}} x_{\mathrm{H}} + \hat{M}_{\mathrm{JJ}} v_{\mathrm{J}}, \tag{8.42}$$

$$x_{\mathrm{H}} \geq 0, \ x_{\mathrm{J}} \geq 0, \ v_{\mathrm{H}} \geq 0, \ v_{\mathrm{J}} \geq 0, \tag{8.43}$$

$$x_{\mathrm{H}}' v_{\mathrm{H}} = x_{\mathrm{J}}' v_{\mathrm{J}} = 0. \tag{8.44}$$

It is important to remember that the index set J *represents the indices of the currently basic* x*-variables, and* H *is the complement of* J *with respect to* N. These two index sets need not have the cardinalities they originally had in (8.18).

The vector $(\hat{q}_{\mathrm{H}}, \hat{q}_{\mathrm{J}})$ is nonnegative as it comes from the final tableau gotten by solving the problem with $\lambda = 0$. For all larger values of $\lambda \le \lambda_{\max}$, the goal is to maintain the nonnegativity of the representation of $r(\lambda) = (\hat{q}_{\mathrm{H}} + \lambda \hat{d}_{\mathrm{H}}, \hat{q}_{\mathrm{J}} + \lambda \hat{d}_{\mathrm{J}})$, i.e., to express it as an element of a complementary cone. If $(\hat{d}_{\mathrm{H}}, \hat{d}_{\mathrm{J}}) \ge (0, 0)$, then the half-line remains in the current complementary cone and λ can be increased all the way to $+\infty$ without ever encountering a boundary of the cone. If $(\hat{d}_{\mathrm{H}}, \hat{d}_{\mathrm{J}})$ has at least one negative component, then to preserve the nonnegativity of the basic variables, we must have

$$\lambda \le \min_{i \in \mathrm{N}} \left\{ \frac{-\hat{q}_i}{\hat{d}_i} : \hat{d}_i < 0 \right\}. \tag{8.45}$$

This is just a standard minimum ratio test. By the nondegeneracy assumption, there is just one index, say $i = t$ for which the minimum ratio is attained. Let

$$\lambda_1 = \min_{i \in \mathrm{N}} \left\{ \frac{-\hat{q}_i}{\hat{d}_i} : \hat{d}_i < 0 \right\} = -\frac{\hat{q}_t}{\hat{d}_t}. \tag{8.46}$$

When $\lambda = \lambda_1$ the corresponding point on the half-line $\mathcal{H}$ belongs to the facet of the cone spanned by the N $-$ 1 currently basic columns whose index is not t. The two vectors corresponding to the columns with index t lie on opposite sides of the subspace $\mathcal{S}$ spanned by the generators of the aforementioned facet.

With $\lambda_1 = -\bar{q}_t/\bar{d}_t$ as defined above, if $\lambda = \lambda_1 + \varepsilon$ where $\varepsilon > 0$ is sufficiently small, the t-th basic variable will be negative. If $t \in$ H, then the basic variable v_t will be negative. The corresponding In-Pivot will express the moving point $r(\lambda)$ for $\lambda = \lambda_1 + \varepsilon$. If $t \in$ J, then the corresponding Out-Pivot will bring about such an expression. The parameter value λ_1 is termed a *breakpoint*.

For $0 \le \lambda < \lambda_1$, we have

$$x_{\mathrm{H}}(\lambda) = 0 \text{ (nonbasic)} \quad \text{and} \quad x_{\mathrm{J}}(\lambda) = \hat{q}_{\mathrm{J}} + \lambda \hat{d}_{\mathrm{J}} \text{ (basic)}. \tag{8.47}$$

The same formula for x is valid when $\lambda = \lambda_1 < \lambda_{\max}$, except that it will now hold for more than one (H, J) pair, reflecting the effect of the pivoting that must occur when λ slightly exceeds the breakpoint λ_1. This process is repeated with the first breakpoint λ_1 playing the role previously played by $\lambda_0 = 0$.

Under the assumption that the current breakpoint $\lambda_\ell < \lambda_{\max}$, we follow the procedure by which $\lambda_{\ell-1}$ was found. This generates a (finite) sequence of breakpoints $\lambda_1 < \lambda_2 < \cdots < \lambda_*$ where λ_* denotes the last of these breakpoints. If $\lambda_{\max} < +\infty$, then it is a breakpoint ($\lambda_{\max} = \lambda_*$), but there is no feasible basis for the system with a larger λ. The case where all the coefficients of λ are nonnegative reveals that $\lambda_{\max} = +\infty$, then $\lambda_* < \lambda_{\max}$ and there is just one set of basic variables for $\lambda \ge \lambda_*$.

Starting from $\boldsymbol{\lambda = \lambda_{\max} = \infty}$. Under favorable conditions (such as those implied by (A4) and (A5)), it is possible to initiate the parametric process by decreasing λ from an extremely large value. This is the approach used by Markowitz in his CLA. Justification for this rests on the following lemma.

Lemma 8.2. *If $A_{\bullet \mathrm{J}}$ is an optimal basis for the linear program*

$$\begin{array}{ll} \textit{minimize} & p'x \\ \textit{subject to} & Ax = b, \\ & x \geq 0, \end{array}$$

and if the reduced costs associated with $A_{\bullet \mathrm{J}}$ are all positive, then pivoting on the nonsingular matrix

$$P = \begin{bmatrix} C_{\mathrm{JJ}} & A'_{\bullet \mathrm{J}} \\ -A_{\bullet \mathrm{J}} & 0 \end{bmatrix}$$

in the KKT system (8.9), (8.10) leads to tableau (8.20) and yields a solution of the KKT system for the PQP (8.1) for all sufficiently large $\lambda > 0$.

Proof. Equations (8.21)–(8.35) describe the effect of such a pivot for an *arbitrary* basis in A. The fact that $A_{\bullet \mathrm{J}}$ is an optimal basis for the given linear program means that $\bar{q}_{\mathrm{J}} = A_{\bullet \mathrm{J}}^{-1} b \geq 0$. The assumed positivity of the reduced costs associated with $A_{\bullet \mathrm{J}}$ implies that

$$\bar{d}_{\mathrm{H}} = d_{\mathrm{H}} - A'_{\bullet \mathrm{H}}(A'_{\bullet \mathrm{J}})^{-1} d_{\mathrm{J}} = p_{\mathrm{H}} - A'_{\bullet \mathrm{H}}(A'_{\bullet \mathrm{J}})^{-1} p_{\mathrm{J}} > 0.$$

As noted in (8.25), we have $\bar{d}_{\mathrm{J}} = 0$. Since the $x_{\mathrm{K}} = u$ is not sign restricted, and $v_{\mathrm{K}} = y = 0$, it now can be seen that we obtain a solution of the KKT conditions for every sufficiently large value of λ, namely

$$\begin{aligned} x_{\mathrm{H}}=0, \quad x_{\mathrm{J}}=A_{\bullet \mathrm{J}}^{-1} b, \quad v_{\mathrm{H}}= & -A'_{\bullet \mathrm{H}}(A'_{\bullet \mathrm{J}})^{-1} C_{\mathrm{JJ}} A_{\bullet \mathrm{J}}^{-1} b + C_{\mathrm{HJ}} A_{\bullet \mathrm{J}}^{-1} b \\ & + \lambda \left(d_{\mathrm{H}} - A'_{\bullet \mathrm{H}}(A'_{\bullet \mathrm{J}})^{-1} d_{\mathrm{J}} \right), \quad v_{\mathrm{J}}=0. \end{aligned}$$

By a "sufficiently large" value of λ we mean $\lambda \geq \max\{0, \max\{-\bar{q}_h/\bar{d}_h : h \in \mathrm{J}\}\}$.□

8.3.3 On Markowitz's Critical Line Algorithm for Quadratic Programming

We caution the reader that we are about to take some notational liberties, which could lead to serious confusion. For the most part, we continue the notation used above which follows that of Wolfe rather than Markowitz. The following table of correspondences may help in cases where there is an important difference.

	WOLFE	MARKOWITZ
Parameter	λ	λ_E
Decision variables	x	X
Coefficients for linear term	p	μ
Lagrange Multipliers on $Ax = b$	u	λ
Lagrange Multipliers on $x \geq 0$	v	η

The CLA for (8.1) begins by finding an optimal basic feasible solution to the linear program (8.2). With assumptions (A4) and (A5) in force, this can be done separately using the simplex method for linear programming.[6] With the nondegeneracy assumption (A6) in force, Lemma 8.2 assures us that with λ sufficiently large, a solution to the KKT conditions for (8.1) is available.

Conceptually, this can be viewed as the initialization step of the CLA. Chapter 13 of (Markowitz, 2000, p. 312) (written by Peter Todd) contains a high-level summary of the CLA. We paraphrase this passage to keep the notation and wording in line with that of a parametric convex programming problem.

> In each iteration of the CLA, we start with a basic feasible solution of the KKT conditions for (8.1), and index sets J and H. (Markowitz denotes these sets IN and OUT, respectively. In general, they are the index sets of the basic and nonbasic x-variables in the current basic feasible solution of the KKT conditions for (8.1).) In the first iteration, these are provided by the simplex algorithm. In subsequent iterations, they are provided by the previous CLA iterations. In each iteration we determine how much we can decrease λ until a variable needs to go IN or OUT, that is, needs to become basic or nonbasic. As λ decreases, we test for the first basic x-variable to cross its lower bound 0, or the first basic v-variable to cross its lower bound 0. The next basic feasible solution of the KKT conditions for (8.1) is the point determined by the basis change corresponding to this test. The critical line algorithm iterates until λ has decreased to its lowest permissible level.

This description of the CLA should prompt visions of a solution process for a PLCP, especially as formulated in the previous subsections. But the pivoting techniques used therein differ from those used by Markowitz. Pivoting can be said to have drawbacks such as error build-up and large computational demands. However, from the conceptual point of view (which can assume exact arithmetic and dismiss the computational effort), pivoting expresses all the data in terms of the current basis and thereby makes it relatively easy to decide whether to terminate the process or to execute another iteration – and, if so, where to pivot again so as to change the basis and re-represent the data. The CLA as developed by Markowitz illustrates the adage that *optimization is all about solving systems of equations; the trick is to figure out which ones to solve.*

In this subsection, we shall discuss what needs to be computed in the CLA and how it can be done. Along the way, we look for their counterparts in the principal pivoting method and show, thereby, that the two algorithms produce the same iterates.

---

[6] We note that without assuming (A5), the boundedness of the feasible region $\mathcal{X}$, an optimal solution of the LP (8.2) may fail to exist. This points to a difference between Markowitz's CLA and Wolfe's SMQP as applied to general convex quadratic programming problems. The latter can be started at $\lambda = 0$.

Suppose the solution of the LP (8.2) terminates with $A_{\bullet \mathrm{J}}$ as an optimal basis in A. In Lemma 8.1 we found that the matrix

$$P = \begin{bmatrix} C_{\mathrm{JJ}} & A'_{\bullet \mathrm{J}} \\ -A_{\bullet \mathrm{J}} & 0 \end{bmatrix}$$

is nonsingular.[7] For ease of reference, we restate that

$$P^{-1} = \begin{bmatrix} 0 & -A_{\bullet \mathrm{J}}^{-1} \\ (A'_{\bullet \mathrm{J}})^{-1} & (A'_{\bullet \mathrm{J}})^{-1} C_{\mathrm{JJ}} A_{\bullet \mathrm{J}}^{-1} \end{bmatrix}.$$

In Lemma 8.2, we showed (by pivoting on P in the KKT system) that from such a matrix $A_{\bullet \mathrm{J}}$ we can obtain a solution of the KKT conditions for (8.1) for sufficiently large λ.

The first question is: How much can λ be reduced before the positivity of some x-variable or some v-variable is lost? Because the CLA does not maintain a tabular form from which one can easily read off the values of the variables, there is some equation-solving to be done. We explain this now; in the process, we relate the steps to those of the parametric principal pivoting method and the theory of the linear complementarity problem.

Using the convenient vector y, we write the equations of the KKT system as

$$\begin{array}{rcl} Cx + A'u - v & = & -\lambda p \\ -Ax \qquad\qquad - y & = & -b \end{array} \tag{8.48}$$

The optimal solution of the LP gives us the index set J of the basic variables. From this we infer the complementary index set $\mathrm{H} = \mathrm{N} \setminus \mathrm{J}$. The system (8.48) can be written in more detail as

$$\begin{bmatrix} C_{\mathrm{HH}} & C_{\mathrm{HJ}} & A'_{\bullet \mathrm{H}} & -I & 0 & 0 \\ C_{\mathrm{JH}} & C_{\mathrm{JJ}} & A'_{\bullet \mathrm{J}} & 0 & -I & 0 \\ -A_{\bullet \mathrm{H}} & -A_{\bullet \mathrm{J}} & 0 & 0 & 0 & -I \end{bmatrix} \begin{bmatrix} x_{\mathrm{H}} \\ x_{\mathrm{J}} \\ u \\ v_{\mathrm{H}} \\ v_{\mathrm{J}} \\ y \end{bmatrix} = - \begin{bmatrix} 0 \\ 0 \\ b \end{bmatrix} - \lambda \begin{bmatrix} p_{\mathrm{H}} \\ p_{\mathrm{J}} \\ 0 \end{bmatrix}. \tag{8.49}$$

[7] Clearly so is the matrix

$$\begin{bmatrix} C_{\mathrm{JJ}} & A'_{\bullet \mathrm{J}} \\ A_{\bullet \mathrm{J}} & 0 \end{bmatrix},$$

but for the time being, we stick with P, even though this is not the way Markowitz puts it. Ultimately, the two are equivalent.

Let D denote the $(n+m) \times 2(n+m)$ coefficient matrix on the left-hand side of (8.49). The blocks shown in D have been arranged to facilitate the presentation. Writing D this way may very well call for the permutation of rows and columns of the (analogous but not shown) coefficient matrix in (8.48). For example, if $m = 2,\ n = 5$, and $\mathrm{J} = \{2, 5\}$, the matrix D has the form

$$D = \begin{bmatrix} c_{11} & c_{13} & c_{14} & c_{12} & c_{15} & a_{11} & a_{21} & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ c_{31} & c_{33} & c_{34} & c_{32} & c_{35} & a_{13} & a_{23} & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ c_{41} & c_{43} & c_{44} & c_{42} & c_{45} & a_{14} & a_{24} & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ c_{21} & c_{23} & c_{24} & c_{22} & c_{25} & a_{12} & a_{22} & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ c_{51} & c_{53} & c_{54} & c_{52} & c_{55} & a_{15} & a_{25} & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ -a_{11} & -a_{13} & -a_{14} & -a_{12} & -a_{15} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ -a_{21} & -a_{23} & -a_{24} & -a_{22} & -a_{25} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}.$$

The (complementary) submatrix

$$B = \begin{bmatrix} -I & C_{\mathrm{HJ}} & A'_{\bullet\mathrm{H}} \\ 0 & C_{\mathrm{JJ}} & A'_{\bullet\mathrm{J}} \\ 0 & -A_{\bullet\mathrm{J}} & 0 \end{bmatrix} \tag{8.50}$$

is clearly invertible. Indeed,

$$B^{-1} = \begin{bmatrix} -I & A'_{\bullet\mathrm{H}}(A'_{\bullet\mathrm{J}})^{-1} & -C_{\mathrm{HJ}}A_{\bullet\mathrm{J}}^{-1} + A'_{\bullet\mathrm{H}}(A'_{\bullet\mathrm{J}})^{-1}C_{\mathrm{JJ}}A_{\bullet\mathrm{J}}^{-1} \\ 0 & 0 & -A_{\bullet\mathrm{J}}^{-1} \\ 0 & (A'_{\bullet\mathrm{J}})^{-1} & (A'_{\bullet\mathrm{J}})^{-1}C_{\mathrm{JJ}}A_{\bullet\mathrm{J}}^{-1} \end{bmatrix}.$$

The basic variables in the solution of the KKT conditions are gotten by solving the system

$$\begin{bmatrix} -I & C_{\mathrm{HJ}} & A'_{\bullet\mathrm{H}} \\ 0 & C_{\mathrm{JJ}} & A'_{\bullet\mathrm{J}} \\ 0 & -A_{\bullet\mathrm{J}} & 0 \end{bmatrix} \begin{bmatrix} v_{\mathrm{H}} \\ x_{\mathrm{J}} \\ u \end{bmatrix} = -\begin{bmatrix} 0 \\ 0 \\ b \end{bmatrix} - \lambda \begin{bmatrix} p_{\mathrm{H}} \\ p_{\mathrm{J}} \\ 0 \end{bmatrix}. \tag{8.51}$$

It is evident from (8.51) that the values of $v_{\mathrm{H}},\ x_{\mathrm{J}},$ and u are (piecewise) linear functions of λ. Relying on (A6), we know that for sufficiently large values of λ, the vectors v_{H} and x_{J} will be positive.

To determine how small λ can become before this positivity ceases to hold, we need to find the solutions to the equations

$$B\bar{q} = q \quad \text{and} \quad B\bar{d} = d,$$

where $q = (0, 0, b)$ and $d = (p_{\mathrm{H}}, p_{\mathrm{J}}, 0)$. These, of course, are

$$\bar{q} = B^{-1}q \quad \text{and} \quad \bar{d} = B^{-1}d.$$

It is clear that because λ is to be *decreased*, the *positive* components of $\bar{d}$ are the ones that matter. For all $i \in N$, we wish to maintain the condition $\bar{q}_i + \lambda \bar{d}_i \geq 0$, so we must have

$$\lambda \geq \max \left\{ 0, \max_i \left\{ -\frac{\bar{q}_i}{\bar{d}_i} : \bar{d}_i > 0 \right\} \right\}.$$

When $\max_i \{ -\frac{\bar{q}_i}{\bar{d}_i} : \bar{d}_i > 0 \} > 0$ there will be a unique index t such that

$$t = \arg\max \left\{ -\frac{\bar{q}_i}{\bar{d}_i} : \bar{d}_i > 0 \right\}.$$

If $t \in \mathrm{H}$, then v_t is the first (sign-restricted) basic variable to reach zero; if $t \in \mathrm{J}$, then x_t is the first such variable to reach zero. When $t \in \mathrm{H}$, we must remove v_t from the basic set of variables. To maintain the solution of the KKT system, its place must be taken by x_t. This gives what is called an In-Pivot. It increases the cardinality of J as t is transferred from H to J. In the other case, x_t must be made nonbasic and replaced by a v_t. This gives an Out-Pivot. Whichever case we have ($t \in \mathrm{H}$ or $t \in \mathrm{J}$), we define

$$\lambda_1 = -\frac{\bar{q}_t}{\bar{d}_t}.$$

At this stage, the CLA calls for a basis change. We emphasize that, throughout the CLA, we are dealing with complementary basic solutions of the system (8.49) in which the vectors of basic variables are (generically) v_{H}, x_{J}, and u. The composition of the index set J (and hence H) changes from one iteration to the next. Note that (A6) implies that the cardinality of J is never less than m, for otherwise b would be expressible as a nonnegative linear combination of fewer than m columns of A.

Suppose $t \in \mathrm{H}$. This means x_t is to be made basic in place of v_t, hence an In-Pivot. The column of B associated with v_t is to be replaced by the column of D associated with x_t. It is important to realize that, in putting the complementary basis matrix B in the form given in (8.50), we have permuted the order of the columns so that the index set H precedes the index set J. Accordingly, caution needs to be exercised in building the new complementary basis.

It can be shown that the *new* matrix B created by such a change of columns is actually nonsingular and thus a complementary basis. We shall not prove this here, but comment that the geometric interpretation of the CLA as a PLCP is useful in supplying the right intuition. When Markowitz's critical line[8] meets the boundary of a complementary cone, there must be a full-dimensional complementary cone on the opposite side of the corresponding facet, and that cone is spanned by the columns of the new matrix B, which must therefore be nonsingular.

[8] Strictly speaking, Markowitz's critical line is a half-line because the parameter λ is bounded below.

The analogous revision of the complementary basis matrix B is performed when $t \in$ J. In that case, we have an Out-Pivot. The index sets J and H are revised to account for the replacement of basic variable x_t by v_t.

After the formation of a new complementary basis matrix B, another ratio test is performed based on the updated versions of the vectors q and d. To compute the updates, $\bar{q}$ and $\bar{d}$, one can solve the appropriate equations or else compute them by first finding the inverse of the new basis B, perhaps by means of a rank-one update. See (Dantzig, 1963, pp. 197–200). Once the vectors $\bar{q}$ and $\bar{d}$ are found, the maximum ratio test is performed and a new critical value of λ is obtained.

This process is repeated until the parameter λ cannot be decreased. It is clear that our rendition of the CLA exactly corresponds to the parametric principal pivoting algorithm (PPPA) with three minor exceptions. First, it seem more natural to initiate the PPPA at the low end, $\lambda = 0$, although starting at the other end is possible. Second, in the PPPA, we eliminated the vectors u and y; but this was just an option, the analogous approach could have been followed in the CLA. Third, in the PPPA we used pivoting instead of equation solving or matrix inversion to obtain updated representations of required data. The calculations must lead to the same sequence of trial solutions.

8.3.4 On Wolfe's Simplex Method for Quadratic Programming

Wolfe's paper Wolfe (1959), "The Simplex Method for Quadratic Programming," presents algorithms for solving the problem (8.1) in the case where C is (at least) positive semidefinite. The algorithms obtain a solution of the related KKT conditions (8.4)–(8.7). As discussed above, the existence of vectors u and v satisfying these conditions along with x are *necessary and sufficient* for the optimality of x in (8.1). The first algorithm given is called the "short form"; the second is called the "long form." These algorithms are shown to converge under different conditions. The long form appears the less restrictive of the two, but it is initialized with an application of the short form.

The short form. The short form obtains a solution to (8.4)–(8.7) under one of the following two assumptions:

1. $\lambda = 0$ and C is positive semidefinite
2. C is positive definite and $\lambda \geq 0$ is arbitrary

If (8.1) is feasible, both of these assumptions are sufficient to guarantee that there exists a solution to the corresponding KKT conditions and hence an optimal solution to (8.1) for any fixed $\lambda \geq 0$. Observe that the second of the above assumptions with $\lambda = 1$ corresponds to the case of a *strictly convex* quadratic programming problem.

It is not restrictive to assume $b \geq 0$ and Wolfe does so. Taking (8.4)–(8.6), Wolfe introduces auxiliary vectors of variables denoted w, z^1, and z^2, thereby arriving at the system[9]

$$Cx + A'u - v + z^1 - z^2 \qquad = -\lambda p, \tag{8.52}$$

$$Ax \quad + w = b, \tag{8.53}$$

$$x \geq 0, \quad v \geq 0, \quad w \geq 0, \quad z^1 \geq 0, \quad z^2 \geq 0 \quad u \text{ free}, \tag{8.54}$$

to which he first applies the Simplex Method of linear programming to minimize $\sum_{i=1}^{m} w_i$ while keeping v and u equal to zero. The auxiliary variables provide a starting feasible basis for this LP. If (8.1) is a feasible program and $A \in R^{m\times n}$ has rank m, this process will drive the vector w to zero and identify a set of m columns in A that form a feasible basis for the constraints of (8.1). This amounts to carrying out a Phase One procedure for the constraints of (8.1) but within the system (8.52), (8.53) while maintaining the aforementioned restrictions on v and u.

It is clear that, in a basic solution of the system (8.52), (8.53), not both of the variables z_j^1 and z_j^2 $(j = 1, \ldots, n)$ can be positive as their corresponding columns are linearly dependent. The second part of the short form involves using the Simplex Method of linear programming to reduce the vectors z^1 and z^2 to zero (while keeping $w = 0$ or eliminating that vector and its columns from the problem altogether). The minimization of the sum of the nonzero z-components is carried out under the restriction that not both x_j and v_j may be basic at any time. This ensures that $x'v = 0$ throughout the process. The algorithm converges (finitely) to a solution of the KKT conditions (8.4)–(8.7).

The procedure just outlined has an exact counterpart in the principal pivoting algorithm given earlier in this paper. First, the vector w in (8.53) corresponds to y in (8.8). The algorithms make these vectors nonbasic at value zero; thereafter, they can be removed from the corresponding system. Both algorithms find a feasible basis in A consisting of the columns indexed by $j_1, \ldots, j_m$. Letting $\mathrm{J} = \{j_1, \ldots, j_m\}$, we may denote this feasible basis by $A_{\bullet\mathrm{J}}$. Then – under the assumption that textbook-style pivoting is done – equation (8.53) becomes

$$A_{\bullet\mathrm{J}}^{-1} Ax + A_{\bullet\mathrm{J}}^{-1} w = A_{\bullet\mathrm{J}}^{-1} b \geq 0. \tag{8.55}$$

Let $\mathrm{N} = \{1, \ldots, n\}$ and $\mathrm{H} = \mathrm{N} \setminus \mathrm{J}$. Then $Ax = A_{\bullet\mathrm{H}} x_{\mathrm{H}} + A_{\bullet\mathrm{J}} x_{\mathrm{J}} + w = b$ so that

$$x_{\mathrm{J}} = A_{\bullet\mathrm{J}}^{-1} b - A_{\bullet\mathrm{J}}^{-1} A_{\bullet\mathrm{H}} x_{\mathrm{H}} - A_{\bullet\mathrm{J}}^{-1} w. \tag{8.56}$$

Substituting this expression for x_{J} into (8.52) has the same effect as when the Simplex Method is applied to minimize $\sum_{i=1}^{m} w_i$ subject to (8.52), (8.53), (8.54), while keeping v and u equal to zero. Keeping w nonbasic at value zero preserves the feasibility of x for (8.1).

[9] The order of presentation of the equations given here differs from what Wolfe used.

Using the index sets J and H, we can write (8.52) in greater detail as

$$C_{\mathrm{HH}}x_{\mathrm{H}} + C_{\mathrm{HJ}}x_{\mathrm{J}} + A'_{\bullet\mathrm{H}}u - v_{\mathrm{H}} + z^1_{\mathrm{H}} - z^2_{\mathrm{H}} = -\lambda p_{\mathrm{H}}, \tag{8.57}$$

$$C_{\mathrm{JH}}x_{\mathrm{H}} + C_{\mathrm{JJ}}x_{\mathrm{J}} + A'_{\bullet\mathrm{J}}u - v_{\mathrm{J}} + z^1_{\mathrm{J}} - z^2_{\mathrm{J}} = -\lambda p_{\mathrm{J}} \tag{8.58}$$

and then examine the effect that the just-described substitution has upon (8.52). But before doing so, we continue to describe an alternative to Wolfe's method for solving the short form problem. This involves the idea of omitting the vectors z^1 and z^2 in (8.52), forgetting about the minimization associated with them, and simply allowing v to be "unrestricted" in sign. (This variant of Wolfe's method was advanced by Dantzig (Dantzig (1961), Dantzig (1963)) in his algorithm for the solution of (8.1) with $\lambda = 0$ and C positive semidefinite.) The aim in the algorithm is then to bring about the nonnegativity of v while retaining the nonnegativity of x and the condition $x_j v_j = 0$ for all $j \in \mathrm{N}$.

After the above substitution of x_{J} and the elimination of z^1 and z^2, (8.57) and (8.58) can be written as

$$v_{\mathrm{H}} = C_{\mathrm{HJ}}A^{-1}_{\bullet\mathrm{J}}b + \lambda p_{\mathrm{H}} + (C_{\mathrm{HH}} - C_{\mathrm{HJ}}A^{-1}_{\bullet\mathrm{J}}A_{\bullet\mathrm{H}})x_{\mathrm{H}} - C_{\mathrm{HJ}}A^{-1}_{\bullet\mathrm{J}}w + A'_{\bullet\mathrm{H}}u \tag{8.59}$$

$$v_{\mathrm{J}} = C_{\mathrm{JJ}}A^{-1}_{\bullet\mathrm{J}}b + \lambda p_{\mathrm{J}} + (C_{\mathrm{JH}} - C_{\mathrm{JJ}}A^{-1}_{\bullet\mathrm{J}}A_{\bullet\mathrm{H}})x_{\mathrm{H}} - C_{\mathrm{JJ}}A^{-1}_{\bullet\mathrm{J}}w + A'_{\bullet\mathrm{J}}u \tag{8.60}$$

The coefficient matrix of u in (8.58) is nonsingular. Hence u can be expressed as a vector depending on x_{H}, v_{J}, and w, namely

$$\begin{aligned} u = {} & -(A'_{\bullet\mathrm{J}})^{-1}C_{\mathrm{JJ}}A^{-1}_{\bullet\mathrm{J}}b - \lambda(A'_{\bullet\mathrm{J}})^{-1}p_{\mathrm{J}} + \left[(A'_{\bullet\mathrm{J}})^{-1}C_{\mathrm{JJ}}A^{-1}_{\bullet\mathrm{J}}A_{\bullet\mathrm{H}} - (A'_{\bullet\mathrm{J}})^{-1}C_{\mathrm{JH}}\right]x_{\mathrm{H}}9 \\ & +(A'_{\bullet\mathrm{J}})^{-1}C_{\mathrm{JJ}}A^{-1}_{\bullet\mathrm{J}}w + (A'_{\bullet\mathrm{J}})^{-1}v_{\mathrm{J}}. \end{aligned} \tag{8.61}$$

This expression for u can be substituted into (8.59) to give v_{H} as a function of x_{H}, v_{J}, and w. The resulting equation is

$$\begin{aligned} v_{\mathrm{H}} = {} & C_{\mathrm{HJ}}A^{-1}_{\bullet\mathrm{J}}b - A'_{\bullet\mathrm{H}}(A'_{\bullet\mathrm{J}})^{-1}C_{\mathrm{JJ}}A^{-1}_{\bullet\mathrm{J}}b + \lambda(p_{\mathrm{H}} - A'_{\bullet\mathrm{H}}(A'_{\bullet\mathrm{J}})^{-1}p_{\mathrm{J}}) \\ & + \left[C_{\mathrm{HH}} - C_{\mathrm{HJ}}A^{-1}_{\bullet\mathrm{J}}A_{\bullet\mathrm{H}} + A'_{\bullet\mathrm{H}}(A'_{\bullet\mathrm{J}})^{-1}C_{\mathrm{JJ}}A^{-1}_{\bullet\mathrm{J}}A_{\bullet\mathrm{H}} - A'_{\bullet\mathrm{H}}(A'_{\bullet\mathrm{J}})^{-1}C_{\mathrm{JH}}\right]x_{\mathrm{H}} \\ & + A'_{\bullet\mathrm{H}}(A'_{\bullet\mathrm{J}})^{-1}v_{\mathrm{J}} + \left[A'_{\bullet\mathrm{H}}(A'_{\bullet\mathrm{J}})^{-1}C_{\mathrm{JJ}}A^{-1}_{\bullet\mathrm{J}} - C_{\mathrm{HJ}}A^{-1}_{\bullet\mathrm{J}}\right]w \end{aligned} \tag{8.62}$$

These expressions for the basic variables $v_{\mathrm{H}}, x_{\mathrm{J}},$ and u are quite complicated. Fortunately, some simplification is possible. The vector w is hereafter required to be zero and the vector u is not sign restricted. Accordingly, both of these can now be eliminated from the system. Doing this gives

$$\begin{aligned} v_{\mathrm{H}} = {} & C_{\mathrm{HJ}}A^{-1}_{\bullet\mathrm{J}}b - A'_{\bullet\mathrm{H}}(A'_{\bullet\mathrm{J}})^{-1}C_{\mathrm{JJ}}A^{-1}_{\bullet\mathrm{J}}b + \lambda(p_{\mathrm{H}} - A'_{\bullet\mathrm{H}}(A'_{\bullet\mathrm{J}})^{-1}p_{\mathrm{J}}) + \left[C_{\mathrm{HH}} - C_{\mathrm{HJ}}A^{-1}_{\bullet\mathrm{J}}A_{\bullet\mathrm{H}}\right. \\ & \left. + A'_{\bullet\mathrm{H}}(A'_{\bullet\mathrm{J}})^{-1}C_{\mathrm{JJ}}A^{-1}_{\bullet\mathrm{J}}A_{\bullet\mathrm{H}} - A'_{\bullet\mathrm{H}}(A'_{\bullet\mathrm{J}})^{-1}C_{\mathrm{JH}}\right]x_{\mathrm{H}} + A'_{\bullet\mathrm{H}}(A'_{\bullet\mathrm{J}})^{-1}v_{\mathrm{J}} \end{aligned} \tag{8.63}$$

$$x_{\mathrm{J}} = A^{-1}_{\bullet\mathrm{J}}b - A^{-1}_{\bullet\mathrm{J}}A_{\bullet\mathrm{H}}x_{\mathrm{H}}. \tag{8.64}$$

We now seek a nonnegative solution $(x, v) = (x_{\mathrm{H}}, x_{\mathrm{J}}, v_{\mathrm{H}}, v_{\mathrm{J}})$ to (8.63) and (8.64) satisfying complementarity condition (8.7). As the system presently stands, by letting the (nonbasic) vectors x_{H} and v_{J} equal zero, we have $x_{\mathrm{J}} = A_{\bullet\mathrm{J}}^{-1}b \geq 0$ and $x'v = 0$.

Under the assumptions on the problem (8.1) imposed by Wolfe, a solution must exist, and it must agree with that given by the principal pivoting method.

The long form. This begins with the solution of a short form problem; it then begins to modify the parameter λ by treating it as a variable to be maximized by the Simplex Method of linear programming with a restricted basis entry rule. Wolfe's system (8.52), (8.53) has $n + m$ equations. The short form problem having been solved, the vectors w, z^1, and z^2 will have been reduced to zero and can be deleted from the system. Assume this has been done; the remaining vectors are then x, v, and u. Of the associated $2n + m$ variables, exactly $n + m$ of them will be basic: one each from the pairs $\{x_j, v_j\}$ for each $j \in \mathrm{N} = \{1, \ldots, n\}$ and all of the variables u_i. Accordingly, there is a version of the system that corresponds to one like (8.61), (8.63), and (8.64). As we have already noted, the m equations pertaining to the Lagrange multipliers u_i can be ignored. In the long form, the parametric aspect of the problem is handled by regarding λ as a variable and maximizing it[10] subject to the restriction that for $j = 1, \ldots, n$, not both x_j and v_j are allowed to be basic. The vectors x and v remain nonnegative and orthogonal throughout the process. After λ becomes basic and takes on a positive value, there will be an index ℓ such that both x_ℓ and v_ℓ are *nonbasic*. As it happens, the coefficients $\partial\lambda/\partial x_\ell$ and $\partial\lambda/\partial v_\ell$ of these nonbasic variables will have opposite signs, hence only one of them can be increased so as to increase the basic variable λ. Increasing that nonbasic variable will either cause λ to go to ∞ or cause some basic variable to decrease to zero. In the latter case, a basis change occurs at which point there is a different pair of nonbasic variables having the same index. The index ℓ is defined to be that of the current nonbasic pair. The algorithm repeats in this manner. Between basis changes, λ increases at a constant rate determined by the coefficient of the increasing nonbasic variable. When the nonbasic variable must stop increasing because it is blocked by a decreasing basic variable becoming zero, the basis change brings about further increase in λ at a new constant rate. This process generates a sequence of values of λ, which can be construed as *breakpoints* in the piecewise linear expression of x and v. The meaning of all this is made clear by the interpretation of the problem as a PLCP.

Interpretation as a PLCP. We now concentrate our attention on the system of equations given by (8.63) and (8.64). This system will have been transformed through the action of the principal pivoting method that solved (8.1) with $\lambda = 0$. We put the transformed system into the format

$$v_{\mathrm{H}} = \bar{q}_{\mathrm{H}} + \lambda\bar{d}_{\mathrm{H}} + \bar{M}_{\mathrm{HH}}x_{\mathrm{H}} + \bar{M}_{\mathrm{HJ}}v_{\mathrm{J}}, \tag{8.65}$$

$$x_{\mathrm{J}} = \bar{q}_{\mathrm{J}} + \lambda\bar{d}_{\mathrm{J}} + \bar{M}_{\mathrm{JH}}x_{\mathrm{H}} + \bar{M}_{\mathrm{JJ}}v_{\mathrm{J}}. \tag{8.66}$$

[10] Literally, *minimizing* its negative.

It is evident from these two equations that the basic variables v_{H} and x_{J} depend linearly on the parameter λ.

The problem at hand can be viewed as PLCP. In the present case, the PLCP is: For every $\lambda \geq 0$, find a solution $(x, v) = (x_{\mathrm{H}}, x_{\mathrm{J}}, v_{\mathrm{H}}, v_{\mathrm{J}})$ of (8.65) and (8.66) such that

$$x \geq 0, \quad v \geq 0, \quad \text{and} \quad x'v = 0.$$

The PLCP is a parametric version of the ordinary LCP in which λ is fixed at zero or equivalently $\bar{d} = (\bar{d}_{\mathrm{H}}, \bar{d}_{\mathrm{J}})$ is a zero vector.

The assumption that C is positive semidefinite implies that

$$\bar{M} = \begin{bmatrix} \bar{M}_{\mathrm{HH}} & \bar{M}_{\mathrm{HJ}} \\ \bar{M}_{\mathrm{JH}} & \bar{M}_{\mathrm{JJ}} \end{bmatrix}$$

is also positive semidefinite.

Wolfe's long form algorithm and the parametric phase of the principal pivoting algorithm begin with essentially the same system of equations: (8.65) and (8.66). The main difference between the two algorithms is that in the long form, the increasing variable of the nonbasic pair drives the basic variables including the parameter λ, now acting as a basic variable; in the parametric principal pivoting algorithm, the parameter is made to increase (as if it were a nonbasic variable) and the basic variables then react to this movement. While the long form algorithm is governed by the restricted basis entry rule, the parametric principal pivoting algorithm is designed to keep the half-line generated by the parameter within the complementary range of the matrix $\widehat{M}$ as described above.

8.3.5 The Impact of Markowitz's Critical Line Algorithm

As demonstrated above, Harry Markowitz's critical line method did not attract widespread attention in the formative years of quadratic programming. Nevertheless, reading the original version of Markowitz's algorithm (and the research in Barankin and Dorfman (1955), Barankin and Dorfman (1956), Barankin and Dorfman (1958)) did inspire Philip Wolfe to put forward his simplex method for quadratic programming. So strong was the influence that he could recently say Wolfe (2008) "without Markowitz's paper, my paper on the simplex method for quadratic programming would not have been written."

Wolfe produced an algorithm which, while equivalent to Markowitz's on appropriate problems, did attract a following among those interested in mathematical programming in a general sense. One such person was George Dantzig who developed "a variant of the Wolfe-Markowitz algorithms." Dantzig's algorithm (Dantzig (1961), Dantzig (1963)) applies to a restrictive type of convex quadratic programming problem: our (8.1) with $\lambda = 0$. Dantzig set, as a problem for (then doctoral student) Richard Cottle, the extension of his QP algorithm to the more general case of (8.1) with $\lambda = 1$ and possibly $p \neq 0$. The extension was found, and a manuscript was written; the manuscript was turned in, but subsequently lost, never to be seen again. Coincidentally, the same ideas were independently discovered and published in van de Panne and Whinston (1964a), van de Panne and Whinston (1964b), van de Panne and Whinston (1969). Even so, the experience led to the exploration of symmetric duality theory (Cottle (1963), Dantzig et al. (1965)) and what came to be called complementary pivot theory (or the linear complementarity problem), see Lemke and Howson (1964), Lemke (1965), Dantzig and Cottle (1967), Cottle and Dantzig (1968), Cottle (1968), Cottle et al. (1992).

The inherently *parametric* aspect of the problem (8.1) first studied by Markowitz and then by Wolfe had its influence as well. New forms of parametric quadratic programs were advanced in van de Panne (1975), Best (1996), and Väliaho (1994) for instance. In addition to these, algorithms were proposed in Murty (1971) (see Murty (1988)) and Cottle (1972), the latter in response to a theoretical question from structural engineering professor Giulio Maier. Diverse applications of parametric linear complementarity were identified and discussed in Pang et al. (1979). Pang also published papers Pang (1980a), Pang (1980b) focusing on certain types of portfolio optimization problems. The influence of Markowitz and Wolfe (as well as Maier) can be found in Smith (1978) on elastoplastic analysis.

The publications mentioned here provide ample evidence to support the conclusion that the work of Harry Markowitz significantly impacted the development of quadratic programming and linear complementarity, well beyond the limits of portfolio selection theory.

References

E.W. Barankin and R. Dorfman (1955). Toward quadratic programming, Report to the Logistics Branch, Office of Naval Research.

E.W. Barankin and R. Dorfman (1956). A method for quadratic programming, *Econometrica* 24, 340.

E.W. Barankin and R. Dorfman (1958). On quadratic programming, *University of California Publications in Statistics* 2:13, University of California Press, Berkeley, pp. 285–318.

E.M.L. Beale (1955). On minimizing a convex function subject to linear inequalities, *Journal of the Royal Statistical Society (B)* 17, 173–184.

E.M.L. Beale (1959). On quadratic programming, *Naval Research Logistics Quarterly* 6, 227–243.

E.M.L. Beale (1968). *Mathematical Programming in Practice*. Sir Isaac Pitman and Sons, London.

M.J. Best (1984). Equivalence of some quadratic programming algorithms, *Mathematical Programming* 30, 71–87.

M.J. Best (1996). An algorithm for the solution of the parametric quadratic programming problem, in (H. Fischer, B. Riedmüller and S. Schäffler, eds.) *Applied Mathematics and Parallel Computing—Festschrift for Klaus Ritter*. Physica-Verlag, Heidelberg, pp. 57–76.

J.C.G. Boot (1964). *Quadratic Programming*. Rand McNally and Company, Chicago.

Y-Y. Chang and R.W. Cottle (1980). Least-index resolution of degeneracy in quadratic programming, *Mathematical Programming* 18, 127–137.

R.W. COTTLE (1963). Symmetric dual quadratic programs, *Quarterly of Applied Mathematics* 21, 237–243.

R.W. COTTLE (1964). Note on a fundamental theorem in quadratic programming, *Journal of the Society for Industrial and Applied Mathematics* 12, 663–665.

R.W. COTTLE (1968). The principal pivoting method of quadratic programming, in (G.B. Dantzig and A.F. Veinott, Jr., eds.) *Mathematics of the Decision Sciences, Part 1*. American Mathematical Society, Providence, R.I., pp. 144–162.

R.W. COTTLE (1972). Monotone solutions of the parametric linear complementarity problem, *Mathematical Programming* 3, 210–224.

R.W. COTTLE AND Y-Y. CHANG (1992). Least-index resolution of degeneracy resolution of degeneracy in linear complementarity with sufficient matrices, *SIAM Journal on Matrix Analysis and Applications* 13, 1131–1141.

R.W. COTTLE AND G.B. DANTZIG (1968). Complementary pivot theory of mathematical programming, *Linear Algebra and its Applications* 1, 103–125.

R.W. COTTLE AND A. DJANG (1979). Algorithmic equivalence in quadratic programming I: A least-distance programming problem, *Journal of Optimization Theory and Applications* 28, 275–301.

R.W. COTTLE, G.J. HABETLER, AND C.E. LEMKE (1970). Quadratic forms semi-definite over convex cones, in (H.W. Kuhn, ed.) *Proceedings of the Princeton Symposium on Mathematical Programming*. Princeton University Press, Princeton, N.J., pp. 551–565.

R.W. COTTLE, J.S. PANG, AND R.E. STONE (1992). *The Linear Complementarity Problem*. Academic Press, Boston.

R.W. COTTLE AND R.E. STONE (1983). On the uniqueness of solutions to linear complementarity problems, *Mathematical Programming* 27, 191–213.

G.B. DANTZIG (1961). Quadratic programming: A variant of the Wolfe-Markowitz algorithms, Research Report 2, Operations Research Center, University of California, Berkeley.

G.B. DANTZIG (1963). *Linear Programming and Extensions*. Princeton University Press, Princeton, N.J.

G.B. DANTZIG AND R.W. COTTLE (1967). Positive (semi-)definite programming, in (J. Abadie, ed.) *Nonlinear Programming*. North-Holland Publishing Company, Amsterdam, pp. 55–73.

G.B. DANTZIG, E. EISENBERG, AND R.W. COTTLE (1965). Symmetric dual nonlinear programs, *Pacific Journal of Mathemtics* 15, 809–812.

G.B. DANTZIG, A. ORDEN, AND P. WOLFE (1955). The generalized simplex method for minimizing a linear form under linear inequality restraints, *Pacific Journal of Mathematics* 5, 183–195.

J.B. DENNIS (1959). *Mathematical Programming and Electrical Networks*. John Wiley & Sons, New York.

A. DJANG (1979). *Algorithmic Equivalence in Quadratic Programming*. PhD thesis, Department of Operations Research, Stanford University, Stanford, Calif.

R. DORFMAN (1951). *Application of Linear Programming to the Theory of the Firm*. University of California Press, Berkeley.

M. FRANK AND P. WOLFE (1956). An algorithm for quadratic programming, *Naval Research Logistics Quarterly* 3, 95–110.

D. GOLDFARB (1972). Extensions of Newton's method and simplex methods for quadratic programs, in (F.A. Lootsma, ed.) *Numerical Methods for Numerical Optimization*. Academic Press, New York, pp. 239–254.

R.L. GRAVES (1967). A principal pivoting simplex algorithm for linear and quadratic programming, *Operations Research* 15, 482–494.

R. L. GRAVES AND P. WOLFE, EDS. (1963). *Recent Advances in Mathematical Programming*. McGraw-Hill, New York.

G. HADLEY (1964). *Nonlinear and Dynamic Programming*. Addison-Wesley Publishing Company, Inc., Reading, Mass.

C. HILDRETH (1954). Point estimates of ordinates of concave functions, *Journal of the American Statistical Association* 49, 598–619.

C. HILDRETH (1957). A quadratic programming procedure, *Naval Research Logistics Quarterly* 4, 79–85.

H.S. HOUTHAKKER (1953). La forme des courbes d'Engel, *Cahiers du Séminaire d'Économetrie* 2 1953, 59–66.

H.S. HOUTHAKKER (1959). The capacity method of quadratic programming, *Econometrica* 28, 62–87.

S. KARLIN (1959). *Mathematical Methods and Theory in Games, Programming, and Economics, Volume I.* Addison-Wesley Publshing Company, Inc., Reading, Mass.

W. KARUSH (1939). Minima of functions of several variables with inequalities as side conditions, Masters Thesis, Department of Mathematics, University of Chicago.

E.L. KELLER (1973). The general quadratic programming problem, *Mathematical Programming* 5, 311–337.

H.W. KUHN AND A.W. TUCKER (1951). Nonlinear programming, in (J. Neyman, ed.) *Second Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, pp. 481–492.

H.P. KÜNZI AND W. KRELLE (1966). *Nonlinear Programming*. Blaisdell, Waltham, Mass. [Translation by F. Levin of *Nichtlineare Programmierung*. Springer-Verlag, Berlin, 1962.]

C.E. LEMKE (1965). Bimatrix equilibrium points and mathematical programming, *Management Science* 11, 681–689.

C.E. LEMKE AND J.T. HOWSON, JR. (1964). Equilibrium points of bimatrix games, *Journal of the Society for Industrial and Applied Mathematics* 12, 413–423.

H. MARKOWITZ (1952). Portfolio selection, *The Journal of Finance* 7, 77–91.

H. MARKOWITZ (1955). The optimization of quadratic functions subject to linear constraints, Research Memorandum RM 1438, The RAND Corporation, Santa Monica, 21 February 1955.

H. MARKOWITZ (1956). The optimization of a quadratic function subject to linear constraints, *Naval Research Logistics Quarterly* 3, 111–133.

H. MARKOWITZ (1959). *Portfolio Selection: Efficient Diversification of Investments*. Wiley, New York. [See also second printing (1970) with corrections and addenda, Yale University Press, New Haven, Ct.]

H. MARKOWITZ (1987). *Mean-Variance Analysis in Portfolio Choice and Capital Markets*. Basil Blackwell, Oxford and Cambridge, Mass.

H. MARKOWITZ (1999). "The early history of portfolio theory: 1600–1960", *Financial Analysts Journal* 55, 5–16.

H. MARKOWITZ (2000). *Mean-Variance Analysis in Portfolio Choice and Capital Markets*. Wiley, New York.

H. MARKOWITZ (2002). Efficient portfolios, sparse matrices, and entities: A retrospective, *Operations Research* 50, 154–160.

K.G. MURTY (1971). On the parametric complementarity problem. Engineering Summer Conference Notes, University of Michigan, Ann Arbor.

K.G. MURTY (1988). *Linear Complementarity, Linear and Nonlinear Programming*. Heldermann-Verlag, Berlin.

J-S. PANG (1980a). A new and efficient algorithm for a class of portfolio selection problems, *Operations Research* 28, 754–767.

J-S. PANG (1980b). A parametric linear complementarity technique for optimal portfolio selection with a risk-free asset, *Operations Research* 28, 927–941.

J-S. PANG (1981). An equivalence between two algorithms for quadratic programming, *Mathematical Programming* 20, 152–165.

J.S. PANG, I. KANEKO, AND W.P. HALLMAN (1979). On the solution of some (parametric) linear complementarity problems with applications to portfolio selection, structural engineering and actuarial graduation, *Mathematical Programming* 16, 325–347.

C. VAN DE PANNE (1975). *Methods for Linear and Quadratic Programming*. North-Holland Publishing Company, Amsterdam.

C. VAN DE PANNE AND A.B. WHINSTON (1964a). The simplex and the dual method for quadratic programming, *Operational Research Quarterly* 15, 355-388.

C. van de Panne and A.B. Whinston (1964b). Simplicial methods for quadratic programming, *Naval Research Logistics Quarterly* 11 273–302.

C. van de Panne and A.B. Whinston (1969). The symmetric formulation of the simplex method for quadratic programming, *Econometrica* 37, 507–527.

A. Perold (1984). Large-scale portfolio optimization, *Management Science* 30, 1143–1160.

D.L. Smith (1978). The Wolfe-Markowitz algorithm for nonholonomic elastoplastic analysis, *Engineering Structures* 1, 8–16.

H. Theil and C. van de Panne (1960). Quadratic programming as an extension of classical quadratic maximization, *Management Science* 7, 1–20.

A.W. Tucker (1963). Principal pivotal transforms of square matrices, *SIAM Review* 5, 305.

H. Väliaho (1994). A procedure for the one-parameter linear complementarity problem, *Optimization* 29, 235–256.

P. Wolfe (1959). The simplex method for quadratic programming, *Econometrica* 27, 382–398.

P. Wolfe (1963). Methods of nonlinear programming, in Graves and Wolfe (1963), pp. 67–86.

P. Wolfe (2008). Private communication.

G. Zoutendijk (1960). *Methods of Feasible Directions*. Van Nostrand, New York.

Chapter 9
Ideas in Asset and Asset–Liability Management in the Tradition of H.M. Markowitz

William T. Ziemba

9.1 Introduction

It is a pleasure to discuss my work that has been influenced by Harry M. Markowitz's mean–variance and other works, and the work of other portfolio theory pioneers such as Bill Sharpe and James Tobin. Section 9.2 discusses pure portfolio theory where it is assumed that means, variances, and other parameters are known and there is a static decision horizon. Section 9.3 discusses the advantage of good means in static portfolio problems and indeed in all models. Section 9.4 discusses my stochastic programming approach to asset–liability management. There we assume that there are scenarios on the asset parameters in the assets-only case and asset and liability parameters in asset–liability situations in multiple periods. Section 9.5 discusses Kelly capital growth investment strategies. Section 9.6 discusses transactions costs and price pressures in portfolio management. Section 9.7 discusses great investors many of which use Kelly betting strategies.

9.2 Static Portfolio Theory

In the static portfolio theory case suppose there are n assets, $i = 1, \ldots, n$, with random returns $\xi_1, \ldots, \xi_n$. The return on asset i, namely ξ_i, is the capital appreciation plus dividends in the next investment period such as monthly, quarterly, or yearly or some other time period. The n assets have the distribution $\mathrm{F}(\xi_1, \ldots, \xi_n)$ with known mean vector $\bar{\xi} = (\bar{\xi}_1, \ldots, \bar{\xi}_n)$ and known $n \times n$ variance–covariance

W.T. Ziemba (✉)
Sauder School of Business (Emeritus), UBC, Vancouver, BC, Canada V6T 1Z2,
and
Mathematical Institute, Oxford University, Oxford, UK
ICMA Centre, University of Reading, Whiteknights, Reading, UK
e-mail: ziemba@interchange.ubc.ca

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_9,
© Springer Science+Business Media, LLC 2010

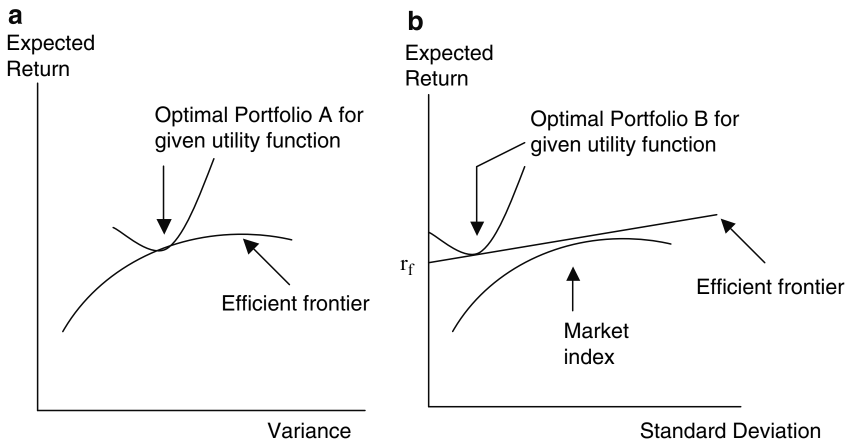
matrix Σ with typical covariance σ_{ij} for $i \neq j$ and variance σ_i^2 for $i = j$. A basic assumption (relaxed in Sect. 9.6) is that the return distributions are independent of the asset weight choices, so $\mathbf{F} \neq \varphi(x)$.

A mean–variance frontier is

$$\begin{aligned} \phi(\delta) = \ & \text{Maximize } \overline{\xi}' x \\ \text{s.t.} \quad & x'\Sigma x \leqslant \delta \\ & e'x = w_0 \\ & x \in K, \end{aligned}$$

where e is a vector of ones, $x = (x_1, \ldots, x_n)$ are the asset weights, K represents other constraints on the x's, and w_0 is the investor's initial wealth.

When variance is parameterized with $\delta > 0$, it yields a concave curve, as in Fig. 9.1a. This is a Markowitz (1952, 1987, 2006) mean–variance efficient frontier and optimally trades off mean which is desirable with variance which is undesirable. Tobin (1958) extended the Markowitz model to include a risk-free asset with mean ξ_0 and no variance. Then the efficient frontier concave curve becomes the straight line as shown in Fig. 9.1b. The standard deviation here is plotted rather than the variance to make the line straight. An investor will pick an optimal portfolio in the Markowitz model using a utility function that trades off mean for variance or, equivalently, standard deviation as shown in Fig. 9.1a to yield portfolio A. For the Tobin model, one does a simpler calculation to find the optimal portfolio which will be on the straight line in Fig. 9.1b between the risk-free asset and the market index M. Here the investor picks portfolio B that is two-thirds cash (risk-free asset) and one-third market index. The market index may be proxied by the S&P500



Markowitz mean-variance efficient frontier

Tobin's risk free asset and separation theorem

Fig. 9.1 Two efficient frontiers

or Wilshire 5000 value weighted indices. Since all investors choose between cash and the market index, this separation of the investor's problem into finding the market index independent of the investor's utility function and then where to be on the line for a given utility function is called Tobin's separation theorem. Ziemba et al. (1974) discuss this and show how to compute the market index and optimal weights of cash and the market index for various utility functions and constraints. They show how to calculate the straight line efficient frontier in Fig. 9.2 with a simple n variable deterministic linear complementary problem (LCP) or a quadratic program. That calculation of Tobin's separation theorem means that for every concave, nondecreasing utility function u the solution of the optimal ratio of risky assets, $i = 1, \ldots, n$ is the same. Hence $x_i^*/x_j^*,\ i, j = 1, \ldots, n,\ i \neq j$ is the same for all $u \in U_2 = [u|u' \geqslant 0,\ u'' \leqslant 0]$.

The portfolio problem is

$$\max_{x \geqslant 0}\ E_{\xi} u(\xi' x), \qquad e'x = 1, \tag{9.1}$$

where $\hat{\xi} = (\xi_0, \xi_1, \ldots, \xi_n) = (\xi_0, \xi)$, ξ_0 is the risk-free asset, and $\xi \sim N(\bar{\xi}, \Sigma)$ and initial wealth $w_0 = 1$.

In step 1, one solves the n-variable deterministic LCP,

$$w = Mz - \bar{\bar{\xi}}, \quad w'z = 0, \quad w \geqslant 0, \quad z \geqslant 0,$$

where the optimal risky asset weights are

$$x_i^* = \frac{z_i^*}{e^* z^*}, \quad i = 1, \ldots, n.$$

This LCP is derived as follows. The slope of the efficient frontier line in Fig. 9.2 is found by

$$\text{Maximizing} \quad \frac{\bar{\xi}' x - \xi_0}{(x' \Sigma x)^{1/2}} \quad \text{s.t.} \quad e'x = 1,\ x \in K,$$

which is equivalent to

$$\text{Max}\, g(x) = \frac{\bar{\bar{\xi}}' x}{(x' \Sigma x)^{1/2}} \quad \text{s.t.} \quad e'x = 1,\ x \in K,$$

where $\bar{\bar{\xi}}_i = \bar{\xi}_i - \xi_0$.

Since $g(x)$ is linear homogeneous and assuming $\lambda x \in K$, where $g(\lambda x) = g(x)$ for $\lambda \geqslant 0$ then the Kuhn–Tucker conditions can be written as the LCP problem above where $w_i = -\bar{\bar{\xi}}_i + \sum_i z$ and $\sum_i$ is the ith row of $\sum$. From the theory of the LCP, see, e.g., Murty (1972), there is a unique solution since $\sum$ is positive definite.

This gives the market index $\mathbf{M} = \sum_{i=1}^{n} \xi_i' x_i^*$ independent of u. Then in step 2, one determines the optimal ratio of $\mathbf{M}$ and the risk-free asset. So for a given $u \in U_2$

from the single variable $\alpha \in (-\infty, 1]$, in Fig. 9.2, $\alpha = 1$ is $\mathbf{M}$, $\alpha = 1/2$ if half $\mathbf{M}$ and half ξ_0, and $\alpha = -3$ means that four units of the risk-free asset ξ_0 is borrowed to invest in $\mathbf{M}$. Then optimal allocation in the risk-free asset is α^* and the optimal allocation in risky asset i, $i = 1, \ldots, n$, is $(1-\alpha^*)x_i^*$, $i = 1, \ldots, n$, which are obtained by solving

$$\max_{-\infty \leqslant \alpha \leqslant 1} E_M u(\alpha\xi_0 + (1-\alpha)M), \tag{9.2}$$

where $M = \xi' x^* \sim N(\bar{\xi}' x^*, x^{*\prime} \Sigma x^*)$ and $\xi = (\xi_1, \ldots, \xi_n)$.

One solves (9.2) using a search algorithm such as golden sections or a bisecting search, see Zangwill (1969). Examples with data and specific utility functions appear in Ziemba et al. (1974). Ziemba (1974) generalized all this to certain classes of infinite variance stable distributions. That analysis, though more complex, follows the normal distribution case. In place of variances, dispersions are used. Under the assumptions one has the convexity of the risk measure property that is analogous to Fig. 9.1. That paper, as well as many other important classic papers, discussions, and extensive problems are reprinted in Ziemba and Vickson (1975, 2006).

Figure 9.2 illustrates the Sharpe (1964)–Lintner (1965)–Mossin (1966) capital asset pricing model which has $E(R_i) = \alpha_i + \beta_i E(R_M)$, where $E(R_i)$ and $E(R_M)$ are the mean return of asset i and the market, respectively, α_i is the excess return from this asset, and β_i is the asset's correlation with the market, M.

Referring to Fig. 9.2, the point A is determined through a tradeoff of expected return and variance. This tradeoff is determined through risk aversion. The standard

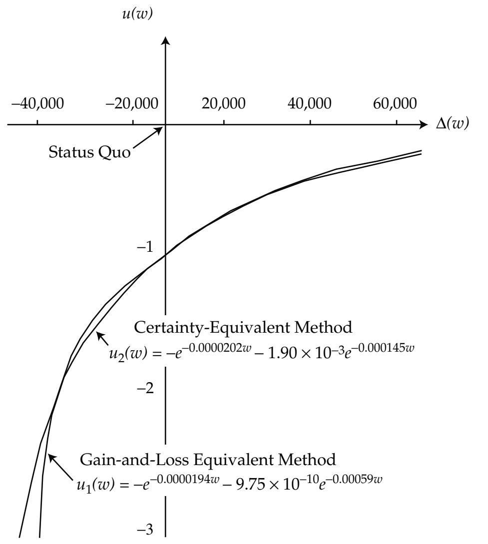


Fig. 9.2 Utility of wealth function for Donald Hausch

way to measure absolute risk aversion is via the Arrow (1965)–Pratt (1964) absolute risk aversion index

$$R_{\mathrm{A}}(w) = \frac{-u''(w)}{u'(w)}.$$

Risk tolerance as used in industry is

$$R_{\mathrm{T}}(w) = \frac{100}{(1/2)R_{\mathrm{A}}(w)}.$$

$R_{\mathrm{A}} = 4$ represents the standard pension fund strategy 60–40 stock bond mix and corresponds to $R_{\mathrm{T}} = 50$ (independent of wealth).

$R_{\mathrm{A}}(w)$ represents all the essential information about the investor's utility function $u(w)$, which is only defined up to a linear transformation, while eliminating everything arbitrary about u, since

$$u(w) \sim \int \mathrm{e}^{\int R_{\mathrm{A}}(w)}.$$

The major theoretical reasons for the use of R_{A} is

$$\begin{aligned} \underbrace{R^1_{\mathrm{A}}(w) \geqslant R^2_{\mathrm{A}}(w)}_{\text{Formula replicates behavior}} &\Longleftrightarrow \pi_1(w,z) \geqslant \pi_2(w,z) \quad \text{for all } w,z \\ &\Longleftrightarrow \underbrace{p_1(w,h) \geqslant p_2(w,h)}_{\text{Behavior}} \quad \text{for all } w,h \end{aligned}$$

where the risk premium $\pi(w,z)$ is defined so that the decision maker is indifferent to receiving the random risk z and the nonrandom amount $\bar{z} - \pi = -\pi$, where $\bar{z} = 0$ and $u[w - \pi] = E_z u(w + z)$ (this defines π). When u is concave, $\pi \geqslant 0$.

For σ_z^2 small, a Taylor series approximation yields $\pi(w,z) = \frac{1}{2}\sigma_z^2 R_{\mathrm{A}}(w) +$ small errors.

For the special risk

$$z = \begin{cases} +h & pr = \frac{1}{2} \\ -h & pr = \frac{1}{2}, \end{cases}$$

where $p(w,h) =$ probability premium $P(z+h) - P(z-h)$ such that the decision maker is indifferent between status quo and the risk z.

$$u(w) = E_z u(w+z) = \frac{1}{2}[1 + p(w,h)]u(w+h) + \frac{1}{2}[1 - p(w,h)]u(x-h). \tag{9.3}$$

Equation (9.3) defines p.

By a Taylor series approximation, $p(w,h) = \frac{1}{2}hR_{\mathrm{A}}(w) +$ small errors.

So risk premiums and probability premiums rise when risk aversion rises and vice versa.

Good utility functions such as those in the general class of decreasing risk aversion functions $u'(w) = (w^a + b)^{-c},\ a > 0,\ c > 0$, which contain log, power, negative power, arctan, etc. Decreasing absolute risk aversion utility functions are preferred because they are the class where a decision maker attaches a positive risk premium to any risk ($\pi(w,z) > 0$), but a smaller premium to any given risk the larger his wealth

$$\left(\frac{\partial \pi(w,z)}{\partial w} \leqslant 0\right).$$

A key result is

$$\underbrace{\frac{\partial R_{\mathrm{A}}(w)}{\partial w} \leqslant 0}_{\text{Replicates behavior}} \Longleftrightarrow \underbrace{\frac{\partial \pi(w,z)}{\partial w} \leqslant 0 \Longleftrightarrow \frac{\partial p(w,h)}{\partial w} \leqslant 0}_{\text{Behavior}} \quad \text{for all } w,z,h > 0.$$

So decreasing absolute risk aversion utility functions replicate those where the risk premium and probability premium are decreasing in wealth.

An individual's utility function and risk aversion index can be estimated using the double exponential utility function

$$u(w) = -\mathrm{e}^{-aw} - b\,\mathrm{e}^{-cw},$$

where a, b, and c are constants. An example using the certainty-equivalent and gain-and-loss equivalent methods is in Fig. 9.2 for Donald Hausch (my co-author of various horseracing books and articles) when he was a student. See the Appendix which describes these estimation methods. This function is strictly concave and strictly increasing and has decreasing absolute risk aversion.

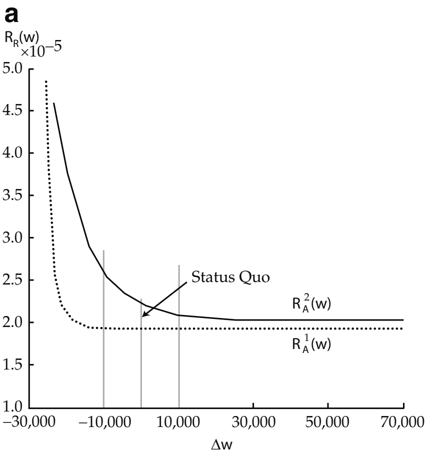
Donald Hausch's utility function was fit by least squares and both methods provide similar curves. His absolute risk aversion is decreasing, see Fig. 9.3a, b, and nearly constant in his investment range where initial wealth w_0 changes by $\pm 10,000$ corresponding to R_{A} between 2.0 and 2.5

$$-w\left[\frac{u''(w)}{u'(w)}\right].$$

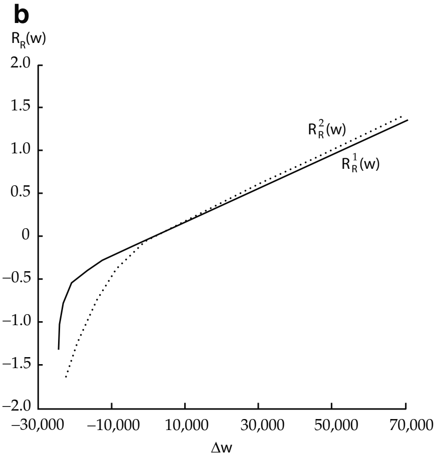
His relative risk aversion is increasing and linear in his investment range.

The Arrow–Pratt risk aversion index is the standard way to measure risk aversion but the little known Rubinstein risk aversion measure is actually optimal. Indeed Kallberg and Ziemba (1983) showed that given utility functions u_1 and u_2, initial wealth w_1 and w_2, and assets $(\xi_1,\ldots,\xi_n) \sim N(\bar{\xi},\Sigma)$. Then if

$$-w_1\frac{E_\xi u_1''(w_1\xi' x^*)}{E_\xi u_1'(w_1\xi' x^*)} = w_2\frac{E_\xi u_2''(w_2\xi' x^*)}{E_\xi u_2'(w_2\xi' x^*)},$$



Absolute risk aversion function estimated by the certainly equivalent method (1) and gain and loss equivalent method (2)



Relative risk aversion function estimated by the certainly equivalent method (1) and gain and loss equivalent method (2)

Fig. 9.3 Risk aversion functions for Donald Hausch

where x^* solves

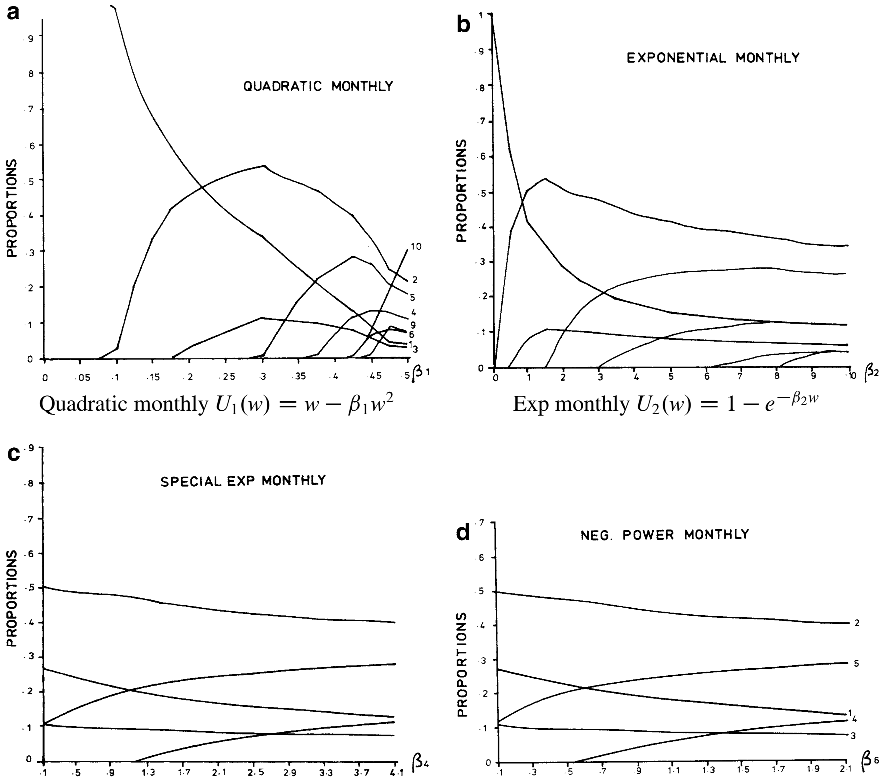
$$\max_{\text{s.t. } x\in K, e'x=w} E_\xi u_1(w_1\xi' x)$$

then x^* solves

$$\max_{\text{s.t. } x\in K, e'x=1} E_\xi u_2(w_2\xi' x).$$

Hence if two investors have the same Rubinstein risk aversion measure, they will then have the same optimal portfolio weights. As shown in Fig. 9.4a–d, special exponential and negative power with the same average risk aversion have very similar mean, variance, expected utility, and portfolio weights and these are very different then with quadratic or exponential utility. When two utility functions have the same average risk aversion, then they have very similar optimal portfolio weights. Here $u_4(w) = \mathrm{e}^{-\beta_4/w}$ and $u_6(w) = (w - w_0)^{-\beta_6}$ are different utility functions but their average risk aversions have similar ranges and these are very different than the quadratic and exponential utilities in Fig. 9.4a, b. The size of the errors is shown in Table 9.1 for fixed risk aversion of $R_{\mathrm{A}} = 4$.

The basic theory of Markowitz, Sharpe, and Tobin assumes that the parameters are known. The real world, of course, does not have constant parameters. So let us look first at what are the errors associated with parameter errors and later at models (stochastic programming) where the probability distribution of asset returns is explicitly considered.



Quadratic monthly $U_1(w) = w - \beta_1 w^2$

Exp monthly $U_2(w) = 1 - e^{-\beta_2 w}$

Special exp monthly $U_4(w) = -e^{\frac{\beta_4}{w}}$

Neg power monthly $U_6(w) = (w - W_0)^{-\beta_6}$

Fig. 9.4 Functional forms asset weights (Source: Kallberg and Ziemba 1983)

9.3 Importance of Means

Means are by far the most important part of any return distribution for actual portfolio results. If you are to have good results in any portfolio problem, you must have good mean estimates for future returns:

> If asset X has cumulative distribution $F(\cdot)$ and asset Y has $G(\cdot)$ and these cumulative distribution functions cross only once, then asset X dominates asset Y for all increasing concave utility functions, that is has higher expected utility, if and only if the mean of X exceeds the mean of Y.

This useful result of Hanoch and Levy (1969) means that the variance and other moments are unimportant for single crossing distributions. Only the means count. With normal distributions X and Y will cross only once if and only if the standard deviation of asset X is less then the standard deviation of asset Y. This is the basic equivalence of Mean–Variance analysis and Expected Utility Analysis via second-order (concave, nondecreasing) stochastic dominance. This is shown in Fig. 9.5 where the second degree and mean–variance dominance is on the left.

Table 9.1 Optimal portfolio weights for alternative utility functions and $\overline{R}_A = 4$ (Source: Kallberg and Ziemba 1983)

Security/ statistic	Exponential (4.0)	Quadratic (0.351447)	Log (–0.832954)	Special exponential (2.884400)	Negative power (1.443557)
1	0.088239	0.082991	0.046975	0.021224	0.047611
2	0.169455	0.165982	0.116220	0.185274	0.112794
3	0.106894	0.106663	0.080160	0.104064	0.079600
4	0.194026	0.198830	0.161247	0.048522	0.154474
5	0.441385	0.445533	0.343318	0.441182	0.328958
6					
7					
8					
9					
10			0.252077	0.199733	0.258232
Mean	1.186170	1.185175	1.151634	1.158397	1.149527
Variance	0.037743	0.037247	0.024382	0.027802	0.023756
Expected utility	0.988236	0.988236	0.987863	0.987589	0.987821
Percent error	—	0	0.703000	0.709900	0.782700

Note: Parameter values are in parentheses. Zeros and blanks indicate values less than 10^{-4}

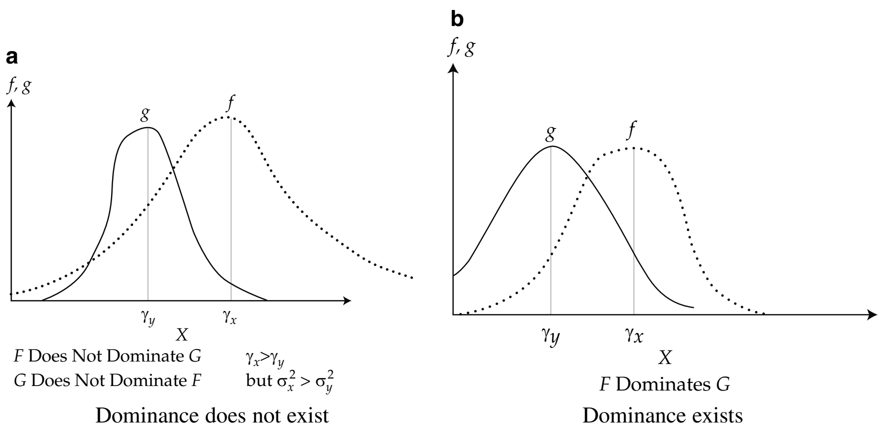


Fig. 9.5 Mean–variance and second-order stochastic dominance

There is no dominance on the right because there are two crosses. This F has a higher mean but also higher variance than G. The densities f and g are plotted here for convenience and give the same results as if the cumulative distribution functions F and G were plotted.

Errors in inputs can lead to significant losses (Fig. 9.6) and larger turnover (Fig. 9.7). Additional calculations appear in Kallberg and Ziemba (1981, 1984) and Chopra and Ziemba (1993).

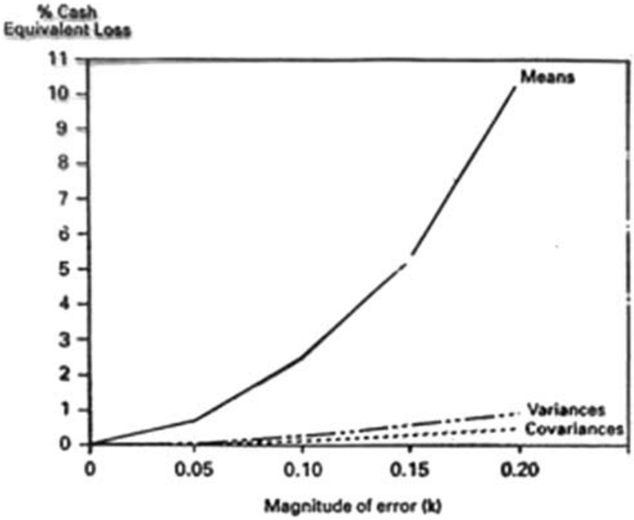


Fig. 9.6 Mean percentage cash equivalent loss due to errors in inputs (Source: Chopra-Ziemba 1993)

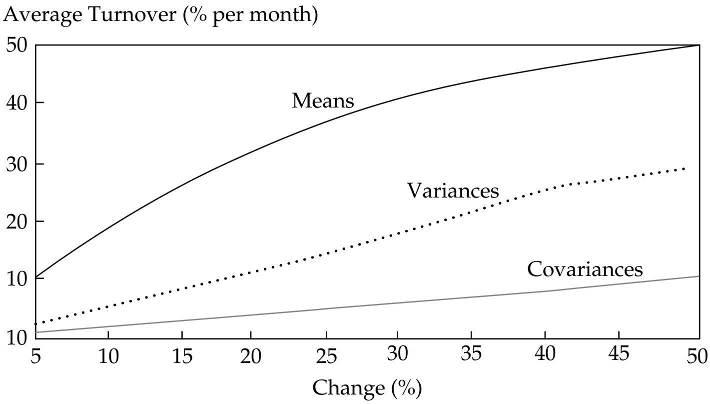


Fig. 9.7 Average turnover for different percentage changes in means, variances, and covariances (Source: Chopra 1993)

The error depends on the risk tolerance, the reciprocal of the Arrow–Pratt risk aversion $\left(R_{\mathrm{T}} = \frac{100}{(1/2)R_{\mathrm{A}}}\right)$. However, errors in means, variances, and covariances are roughly 20:2:1 times as important, respectively. But with low risk aversion, like log, the ratios can be 100:2:1. So good estimates are by far the most crucial aspect for successful application of a mean–variance analysis and we will see that in all other stochastic modeling approaches.

Table 9.2 Average ratio of CEL for errors in means, variances, and covariances (Source: Chopra-Ziemba 1993)

Risk tolerance	Errors in means vs. covariances	Errors in means vs. variances	Errors in variances vs. covariances
25	5.38	3.22	1.67
50	22.50	10.98	2.05
75	56.84	21.42	2.68
	↓	↓	↓
	20	10	2
	Error mean	Error var	Error covar
	20	2	1

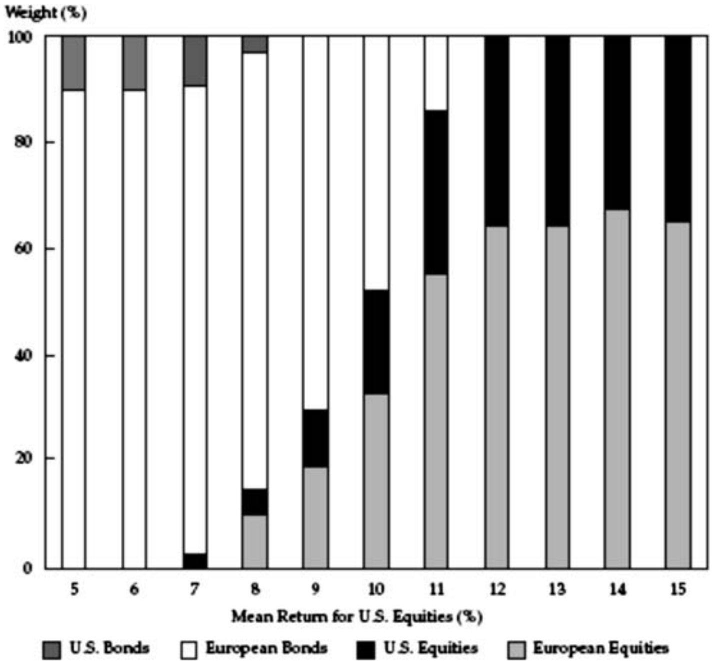
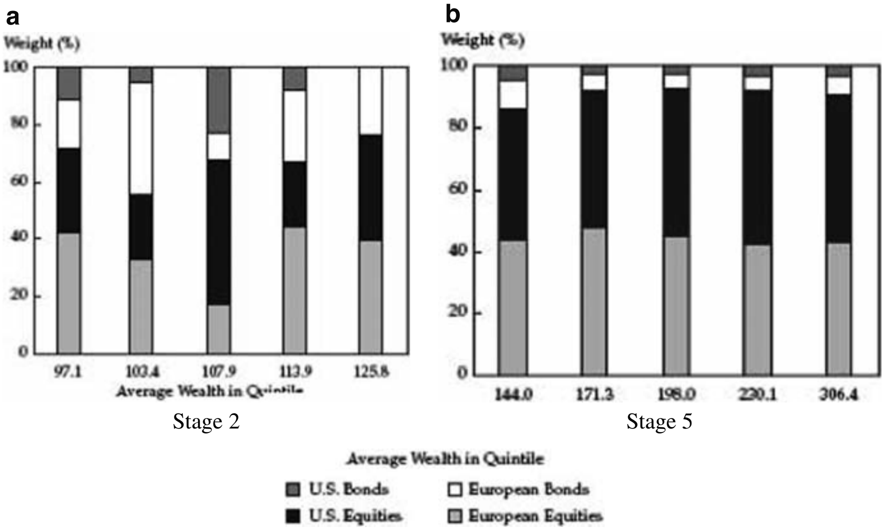


Fig. 9.8 Optimal equity and bond allocations in period 1 of the InnoALM model (Source: Geyer and Ziemba 2008)

The sensitivity of the mean carries into multiperiod models. There the effect is strongest in period 1 then less and less in future periods (see Geyer and Ziemba 2008). This is illustrated in Figs. 9.8 and 9.9 for a five period, 10-year model designed for the Siemen's Austria pension fund that is discussed in Sect. 9.4. There it is seen that in period 1, with bond means for the US and Europe in the 6–7% area, the optimal allocation to European and US equity can be 100% with a mean of 12%+ and about 30% when the mean is 9% or less. Whereas in later periods, this sensitivity is less and, by period 5, it is almost nonexistent.



Stage 2 Stage 5

Fig. 9.9 The effects of state-dependent correlations: optimal weights conditional on quintiles of portfolio weights in periods 2 and 5 of the InnoALM model (Source: Geyer and Ziemba 2008)

9.4 The Stochastic Programming Approach to Asset–Liability Management

I discuss my approach using scenarios and optimization to model asset–liability decisions for pension funds, insurance companies, individuals, retirement, bank trading departments, hedge funds, etc. It includes the essential problem elements: uncertainties, constraints, risks, transactions costs, liquidity, and preferences over time to provide good results in normal times and avoid or limit disaster when extreme scenarios occur. The stochastic programming approach while complex is a practical way to include key problem elements that other approaches are not able to model.

Other approaches (static mean–variance, fixed mix, stochastic control, capital growth, continuous time finance, etc.) are useful for the microanalysis of decisions and the SP approach is useful for the aggregated macro (overall) analysis of relevant decisions and activities. They yield good results most of the time but frequently lead to the recipe for disaster: overbetting and not being truly diversified at a time when an extreme scenario occurs. It pays to make a complex stochastic programming model when a lot is at stake and the essential problem has many complications.

The accuracy of the actual scenarios chosen and their probabilities contributes greatly to model success. However, the scenario approach generally leads to superior investment performance even if there are errors in the estimations of both the actual scenario outcomes and their probabilities. It is not possible to include all scenarios or even some that may actually occur. The modeling effort attempts to cover well in the range of possible future evolution of the economic environment. The predominant view is that such models do not exist, are impossible to successfully implement, or are prohibitively expensive. I argue that better modern computer power, better large-scale stochastic linear programming codes, and better modeling skills such models can be widely used in many applications and are very cost effective.

We know from Harry Markowitz's work and that of others that mean–variance models are useful in an assets-only situation. From my consulting at Frank Russell, I know that professionals adjust means (mean-reversion, James-Stein, etc.) and constrain output weights.

They do not change asset positions unless the advantage of the change is significant. Also, they do not use mean–variance analysis with liabilities and other major market imperfections except as a first test analysis. Mean–variance models can be modified and applied in various ways to analyze many problems, see Markowitz (1987) and Grinold and Khan (1999) for example. Masters of this have been Barra who analyze any question via a simple mean–variance model. But mean–variance models define risk as a terminal wealth surprise regardless of direction and make no allowance for skewness preference. Moreover, they treat assets with option features inappropriately as shown in Fig. 9.10.

But we will use in my models a relative of variance, namely the convex risk measure weighted downside target violations and that will avoid such objections.

Asset–liability problems can be for institutions or individuals. My experience is that the latter is much more difficult because one must make models useful for many individuals, people change their assets, goals and preferences, and hide assets, etc. Basic references are Ziemba and Mulvey (1998) and Zenios and Ziemba (2006, 2007) that contain theory, computations, and many case studies.

Possible approaches to model ALM situations include:

- Simulation: Too much output to understand but very useful as check
- Mean–variance: OK for one period but with constraints, etc.
- Expected log: Very risky strategies that do not diversify well, fractional Kelly with downside constraints are excellent for risky investment betting
- Stochastic control: Bang-bang policies, see the Brennan-Schwartz paper in ZM (1998), how to constrain to be practical?

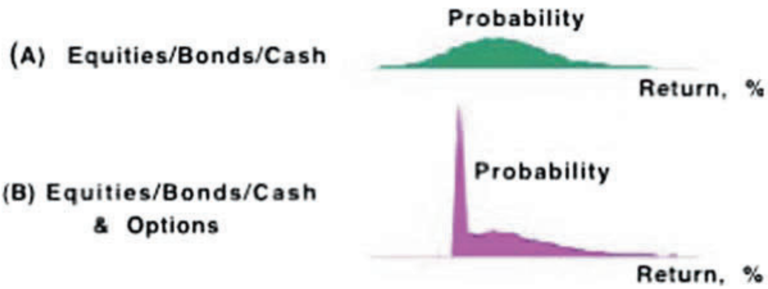


Fig. 9.10 Two distributions with the same mean and variance

- Stochastic programming/stochastic control: Mulvey (see Mulvey 1996, Mulvey et al. 2003, and Mulvey et al. 2007) uses this approach which is a form of volatility pumping, which is discussed in general by Luenberger (1998)
- Stochastic programming: My approach

Continuous time modeling, though popular in the academic world, seems to be impractical, as Fig. 9.12 shows. Here, the asset proportions from a Merton (1992) stochastic control type model are stocks, bonds, and interest rate futures representing spot and future interest rates.

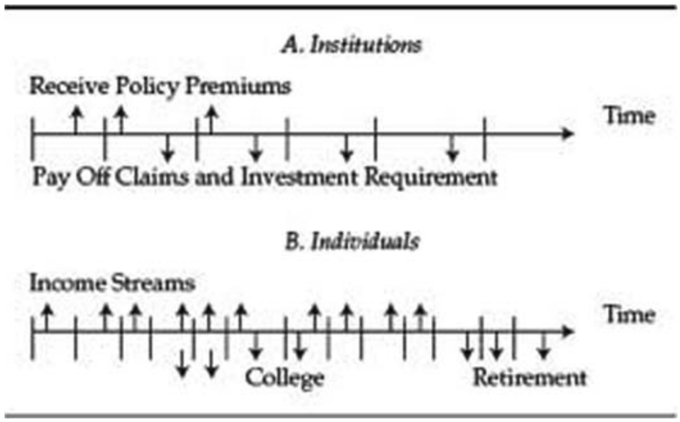


Fig. 9.11 Time and events in institutional and individual ALM models (Source: Ziemba 1993)

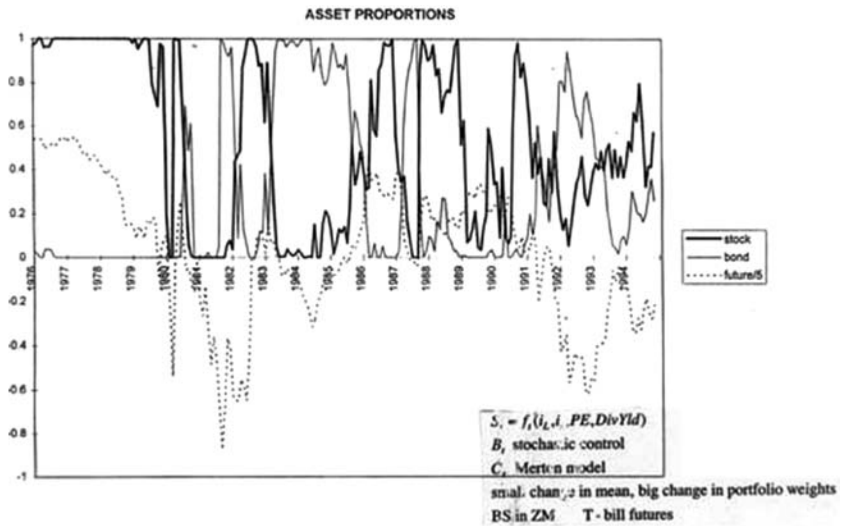


Fig. 9.12 Asset weights over time with a stochastic control continuous time Merton type model (Source: Brennan and Schwartz 1998)

The stochastic programming approach is ideally suited to analyze such problems with the following features:

- Multiple time periods; end effects – steady state after decision horizon adds one more decision period to the model.
- Consistency with economic and financial theory for interest rates, bond prices, etc.
- Discrete scenarios for random elements – returns, liability costs, currency movements.
- Utilize various forecasting models, handle fat tails.
- Institutional, legal, and policy constraints.
- Model derivatives, illiquid assets, and transactions costs.
- Expressions of risk in terms understandable to decision makers; the more you lose, the more is the penalty for losing.
- Maximize long-run expected profits net of expected discounted convex penalty costs for shortfalls; pay more and more penalty for shortfalls as they increase.
- Model as constraints or penalty costs in the objective to maintain adequate reserves and cash levels and meet regularity requirements.
- We can now solve very realistic multiperiod problems on PCs.
- If the current situation has never occurred before, then use one that is similar to add scenarios. For a crisis in Brazil, use Russian crisis data, for example. The results of the SP will give you good advice when times are normal and keep you out of severe trouble when times are bad.
- Those using SP models may lose 5–10–15% but they will not lose 50–70–95% like some investors and hedge funds. If the scenarios are more or less accurate and the problem elements reasonably modeled, the SP will give good advice. You may slightly underperform in normal markets but you will greatly overperform in bad markets when other approaches may blow up.

These models make you diversify which is the key for keeping out of trouble. My work in stochastic programming began in my PhD work at Berkeley. In 1974, I taught my first stochastic programming class and that led to early papers with my students Jarl Kallberg and Martin Kusy (see Kallberg et al. 1982 and Kusy and Ziemba 1986). Table 9.3 illustrates the models we build at Frank Russell in the 1990s.

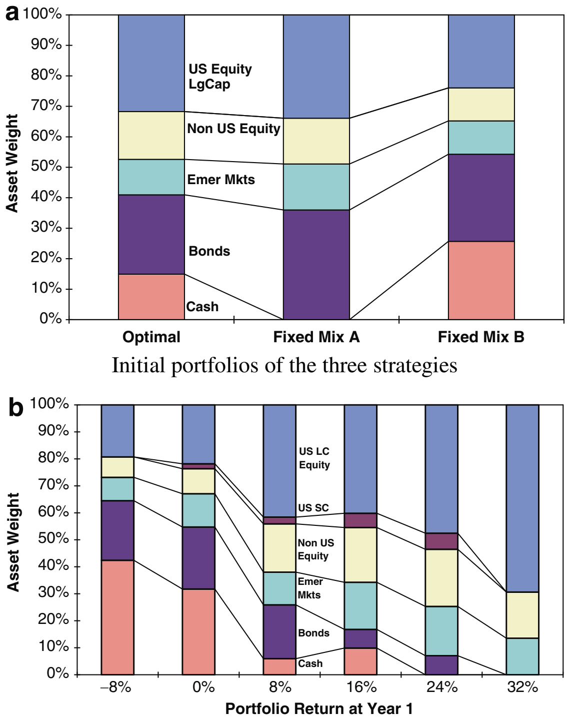
In Cariño et al. (1998), we showed that stochastic programming models usually beat fix-mix models. The latter are basically volatility pumping models (buy low, sell high) (see Luenberger 1998). Figure 9.13 illustrates how the optimal weights change depending on previous periods' results. Despite good results, fixed mix and buy and hold strategies do not utilize new information from return occurrences in their construction. By making the strategy scenario dependent using a multiperiod stochastic programming model, a better outcome is possible.

We compare two strategies as follows:

1. The dynamic stochastic programming strategy which is the full optimization of the multiperiod model.

Table 9.3 Russell business engineering models

Model	Type of application	Year delivered	Number of scenarios	Computer hardware
Russell–Yasuda (Tokyo)	Property and casualty insurance	1991	256	IBM RISC 6000
Mitsubishi trust (Tokyo)	Pension consulting	1994	2,000	IBM RISC 6000 with four Parallel Processers
Swiss bank corp (Basle)	Pension consulting	1996	8,000	IBM UNIX2
Daido life insurance company (Tokyo)	Life insurance	1997	25,600	IBM PC
Banca Fideuram (Rome)	Assets only personal	1997	10,000	IBM UNIX2 and PC
Consulting clients	Assets only institutional	1998	Various	IBM UNIX2 and PC



Initial portfolios of the three strategies

Contingent allocations at year one

Fig. 9.13 Example portfolios

2. The fixed mix in which the portfolios from the mean–variance frontier have allocations rebalanced back to that mix at each stage; buy when low and sell when high. This is like covered calls which is the opposite of portfolio insurance.

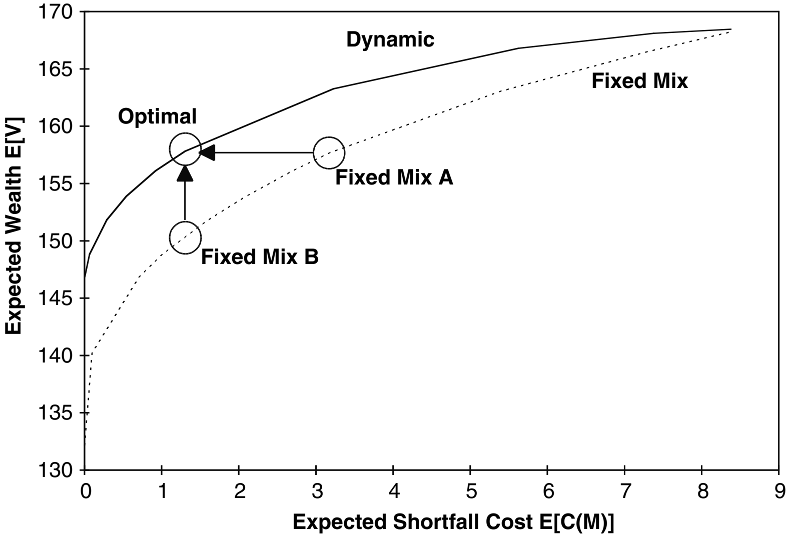


Fig. 9.14 The optimal stochastic strategy dominates fixed mix

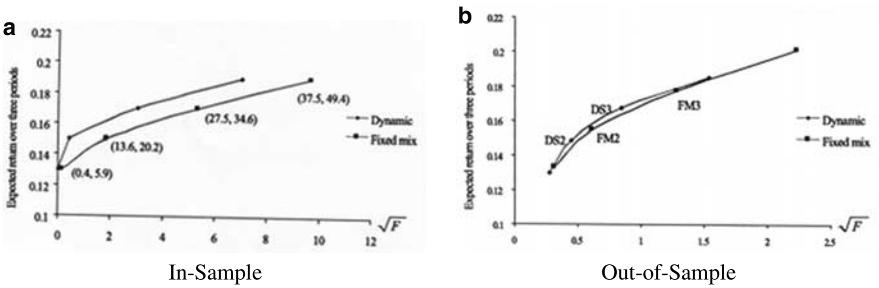


Fig. 9.15 Comparison of advantage of stochastic programming over fixed mix model in and out of sample (Source: Fleten et al. 2002)

Consider fixed mix strategies A (64–36 stock bond mix) and B (46–54 stock bond mix). The optimal stochastic programming strategy dominates as shown in Fig. 9.14.

A further study of the performance of stochastic dynamic and fixed mix portfolio models was made by Fleten et al. (2002). They compared two alternative versions of a portfolio model for the Norwegian life insurance company Gjensidige NOR, namely multistage stochastic linear programming and the fixed mix constant rebalancing study. They found that the multiperiod stochastic programming model dominated the fixed mix approach but the degree of dominance is much smaller out-of-sample than in-sample (see Fig. 9.15). This is because in out-of-sample the random input data is structurally different from in-sample, so the stochastic programming model loses its advantage in optimally adapting to the information available in the scenario tree. Also the performance of the fixed mix approach improves because the asset mix is updated at each stage.

The Russell–Yasuda Kasai was the first large scale multiperiod stochastic programming model implemented for a major financial institution (see Henriques 1991). As a consultant to the Frank Russell Company during 1989–1991, I designed the model. The team of David Cariño, Taka Eguchi, David Myers, Celine Stacy, and Mike Sylvanus at Russell in Tacoma, Washington implemented the model for the Yasuda Fire and Marine Insurance Co., Ltd in Tokyo under the direction of research head Andy Turner. Roger Wets and my former UBC PhD student Chanaka Edirishinghe helped as consultants in Tacoma, and Kats Sawaki was a consultant to Yasuda Kasai in Japan to advise them on our work. Kats, a member of my 1974 UBC class in stochastic programming where we started to work on ALM models, was then a professor at Nanzan University in Nagoya and acted independently of our Tacoma group. Kouji Watanabe headed the group in Tokyo which included Y. Tayama, Y. Yazawa, Y. Ohtani, T. Amaki, I. Harada, M. Harima, T. Morozumi, and N. Ueda.

Experience has shown that we should not be concerned with getting all the scenarios exactly right when using stochastic programming models. You cannot do this and it does not matter much anyway. Rather worry that you have the problems periods laid out reasonably and the scenarios basically cover the means, the tails, and the chance of what could happen.

Back in 1990/1991 computations were a major focus of concern. I knew how to formulate the model, which was an outgrowth of Kallberg et al. (1982) and Kusy and Ziemba (1986). David Carino did much of the formulation details. Originally we had 10 periods and 2,048 scenarios. It was too big to solve at that time and became an intellectual challenge for the stochastic programming community. Bob Entriken, D. Jensen, R. Clark, and Alan King of IBM Research worked on its solution but never quite cracked it. We quickly realized that ten periods made the model far too difficult to solve and also too cumbersome to collect the data and interpret the results and the 2,048 scenarios were at that time a large number to deal with. About 2 years later, Hercules Vladimirou working with Alan King at IBM Research was able to effectively solve the original model using parallel processing on several workstations.

The Russell–Yasuda model was designed to satisfy the following need as articulated by Kunihiko Sasamoto, Director and Deputy President of Yasuda Kasai.

> The liability structure of the property and casualty insurance business has become very complex, and the insurance industry has various restrictions in terms of asset management. We concluded that existing models, such as Markowitz mean variance, would not function well and that we needed to develop a new asset/liability management model.
>
> The Russell–Yasuda Kasai model is now at the core of all asset/liability work for the firm. We can define our risks in concrete terms, rather than through an abstract, in business terms, measure like standard deviation. The model has provided an important side benefit by pushing the technology and efficiency of other models in Yasuda forward to complement it. The model has assisted Yasuda in determining when and how human judgment is best used in the asset/liability process (Carino et al. 1994)

The model was a big success and of great interest both in the academic and institutional investment asset–liability communities.

9.4.1 The Yasuda Fire and Marine Insurance Company

The Yasuda Fire and Marine Insurance Company called Yasuda Kasai meaning fire is based in Tokyo. It began operations in 1888 and was the second largest Japanese property and casualty insurer and seventh largest in the world by revenue. Its main business was voluntary automobile (43.0%), personal accident (14.4%), compulsory automobile (13.7%), fire and allied (14.4%), and other (14.5%). The firm had assets of 3.47 trillion yen (US$26.2 billion) at the end of fiscal 1991 (March 31, 1992). In 1988, Yasuda Kasai and Russell signed an agreement to deliver a dynamic stochastic asset allocation model by April 1, 1991. Work began in September 1989. The goal was to implement a model of Yasuda Kasai's financial planning process to improve their investment and liability payment decisions and their overall risk management. In 1989, after Russell could not figure out how to do this, I proposed a stochastic programming model which we then built at Russell in Tacoma.

The business goals were to:

1. Maximize long-run expected wealth
2. Pay enough on the insurance policies to be competitive in current yield
3. Maintain adequate current and future reserves and cash levels
4. Meet regulatory requirements especially with the increasing number of saving-oriented policies being sold that were generating new types of liabilities

The following is a summary of the Russell–Yasuda model and its results. See also the original papers of Cariño Ziemba et al. (1994, 1998a, b) and the surveys of Ziemba (2006, 2007).

- The model needed to have more realistic definitions of operational risks and business constraints than the return variance used in previous mean–variance models used at Yasuda Kasai.
- The implemented model determines an optimal multiperiod investment strategy that enables decision makers to define risks in tangible operational terms such as cash shortfalls.
- The risk measure used is convex and penalizes target violations more and more, as the violations of various kinds and in various periods increase.
- The objective is to maximize the discounted expected wealth at the horizon net of expected discounted penalty costs incurred during the five periods of the model.
- This objective is similar to a mean–variance model except it is over five periods and only counts risk through downside target violations.
- I greatly prefer this approach to VaR or CVAR and its variants for ALM applications because for most people and organizations, the nonattainment of goals is more and more damaging not linear in the nonattainment (as in CVAR) or not considering the size of the nonattainment at all (as in VaR).
- A reference on VaR and CVAR as risk measures is Artzner et al. (1999). They argue that good risk measures should be coherent and satisfy a set of axioms.

The convex risk measure I use is coherent. Rockafellar and Ziemba (2000) define a set of axioms that justify these convex risk measures.

- The model formulates and meets the complex set of regulations imposed by Japanese insurance laws and practices.
- The most important of the intermediate horizon commitments is the need to produce income sufficiently high to pay the required annual interest in the savings type insurance policies without sacrificing the goal of maximizing long-run expected wealth.
- During the first 2 years of use, fiscal 1991 and 1992, the investment strategy recommended by the model yielded a superior income return of 42 basis points (US$79 million) over what a mean–variance model would have produced. Simulation tests also show the superiority of the stochastic programming scenario-based model over a mean–variance approach.
- In addition to the revenue gains, there have been considerable organizational and informational benefits.
- The model had 256 scenarios over four periods plus a fifth end effects period.
- The model is flexible regarding the time horizon and the length of decision periods, which are multiples of quarters.
- A typical application has initialization, plus period 1 to the end of the first quarter; period 2 the remainder of fiscal year 1; period 3 the entire fiscal year 2; period 4 fiscal years 3, 4, and 5; and period 5, the end effects years 6 on to forever.

Figure 9.16 shows the multistage stochastic linear programming structure of the Russell–Yasuda Kasai model.

The basic formulation of the model is as follows.

The objective of the model is to allocate discounted fund value among asset classes to maximize the discounted expected wealth at the end of the planning horizon T; less expected penalized shortfalls accumulated throughout the planning horizon

$$\text{Maximize } E\left[W_T - \sum_{t=1}^{T} c_i(w_t)\right]$$

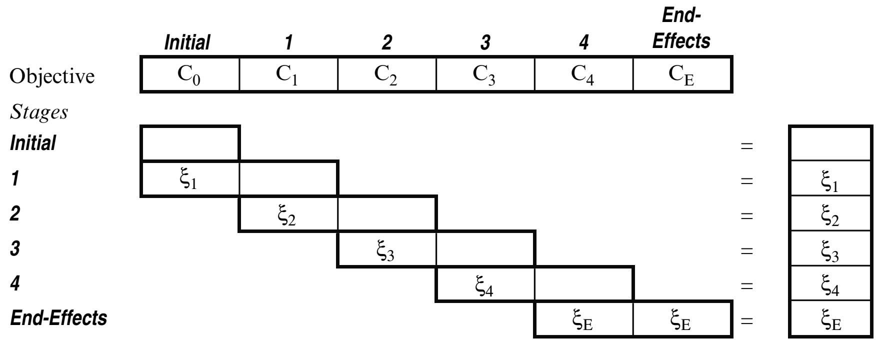


Fig. 9.16 Layout of the Russell–Yasuda Kasai model

subject to budget constraints

$$\sum_n X_{nt} - V_t = 0,$$

asset accumulation relations

$$V_{t+1} - \sum_n (1 + RP_{nt+1} + RI_{nt+1} X_{nt} = F_{t+1} - P_{t+1} - I_{t+1},$$

income shortfall constraints

$$\sum_n RI_{nt+1} X_{nt} + w_{t+1} + (-v_{t+1}) = g_t + L_t,$$

and nonnegativity constraints

$$x_{nt} \geqslant 0, \quad v_{t+1} \geqslant 0, \quad w_{t+1} \geqslant 0,$$

for $t = 0, 1, 2, \ldots, T-1$. Liability balances and cash flows are computed to satisfy the liability accumulation relations

$$L_{t+1} = (1 + g_{t+1})L_t + F_{t+1} - P_{t+1} - I_{t+1}, \quad t = 0, \ldots, T-1.$$

Publicly available codes to solve such models are discussed in Wallace and Ziemba (2005); see also Gonzio and Kouwenberg (2001). The Russell–Yasuda model is small by 2009 standards but in 1991 it was a challenge. The dimensions of a typical implemented problem are shown in Table 9.4.

Figure 9.17 shows Yasuda Kasai's asset–liability decision-making process.

Yasuda Fire and Marine faced the following situation:

- An increasing number of savings-oriented policies were being sold which had new types of liabilities.
- The Japanese Ministry of Finance imposed many restrictions through insurance law that led to complex constraints.
- The firm's goals included both current yield and long-run total return that lead to risks and objectives that were multidimensional.

Table 9.4 The dimensions of a typical implemented problem

Per	BrPar	Scen	Assets	Alloc Var	Rows	Cols	Coeff	GLP: Rows	Cols	Coeff
INI	1	1	7	59	22	60	257	22	60	257
Q01	8	8	7	59	48	85	573	384	680	4,584
Y01	4	32	7	59	54	96	706	1,728	3,072	22,592
Y02	4	128	7	29	52	66	557	6,656	8,448	71,296
Y05	2	256	7	29	52	66	407	13,312	16,896	104,192
YFF	1	256	5	21	35	58	450	8,960	14,848	115,200

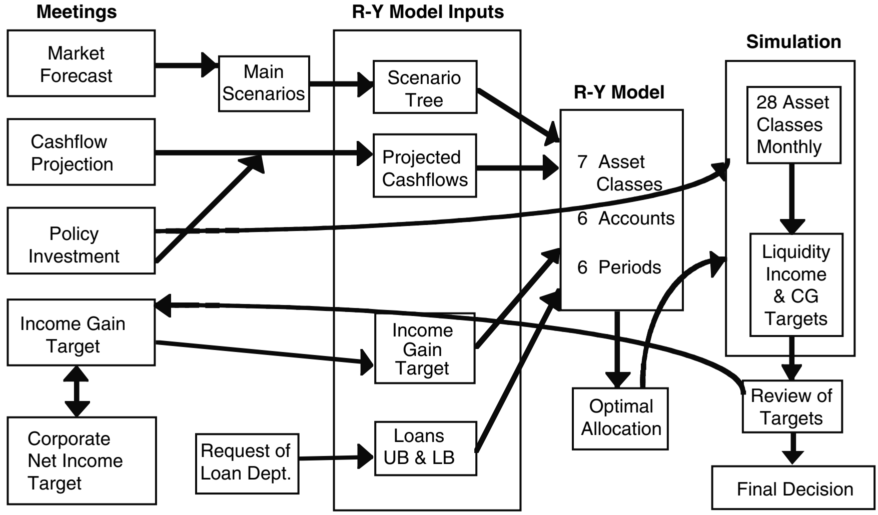


Fig. 9.17 Yasuda Kasai's asset–liability decision process

The insurance policies were complex with a part being actual insurance and another part an investment with a fixed guaranteed amount plus a bonus dependent on general business conditions in the industry. The insurance contracts are of varying length; maturing, being renewed, or starting in various time periods, and they are subject to random returns on assets managed, insurance claims paid, and bonus payments made. There are many regulations on assets including restrictions on equity, loans, real estate, foreign investment by account, foreign subsidiaries, and tokkin (pooled accounts). The asset classes were as follows:

Asset	Associated index
Cash bonds	Nomura bond performance index
Convertible bonds	Nikko research convertible bond index
Domestic equities	TOPIX
Hedged foreign bonds	Salomon Brothers world bond index (or hedged equivalent)
Hedged foreign equities	Morgan Stanley world equity index (or hedged equivalent)
Unhedged foreign bonds	Salomon Brothers world bond index
Unhedged foreign equities	Morgan Stanley world equity index
Loans	Average lending rates (trust/long-term credit (or long-term prime rates)
Money trusts, etc.	Call rates (overnight with collateral)
Life insurance company general accounts	

To divide the universe of available investments into a manageable number of asset classes involves a tradeoff between detail and complexity. A large number

Table 9.5 Expected allocations for initialization period: INI (100 million yen: percentages by account)

	Total	%
Cash	2,053	9
Loans-f1	5,598	26
Loans-fx	5,674	26
Bonds	2,898	13
Equity	1,426	7
Foreign bonds	3,277	15
Foreign equity	875	4
Total	21,800	100

Total book value 1: 22,510
Total book value 2: 34,875

Table 9.6 Expected allocations in the end effects period

	General	Savings	Spec savings 1	Spec savings 2	Exogenous	Total	%
Cash	0	44	0	36	0	80	0.1
Bonds	5,945	17	14,846	1,311	0	22,119	40.1
Equity	0	0	4	0	18,588	18,592	33.7
For bonds	2,837	1,094	0	0	0	3,931	7.1
For equity	0	4,650	6,022	562	0	11,234	20.4
Total	8,782	5,804	20,072	1,908	18,588	55,154	

Total book value 1: 28,566
Total book value 2: 50,547

of asset classes would increase detail at the cost of increasing size. Therefore, the model allows the number and definition of asset classes to be specified by the user. There can be different asset classes in different periods. For example, asset classes in earlier periods can be collapsed into aggregate classes in later periods.

A major part of the information from the model is in the terms of reports consisting of tables and figures of model output. Actual asset allocation results from the model that are confidential. But we have the following expected allocations in the initialization (Table 9.5) and end effects periods (Table 9.6). These are averages across scenarios in 100 million yen units and percentages by account.
In summary:

1. The 1991 Russell Yasuda Kasai Model was then the largest application of stochastic programming in financial services.
2. There was a significant ongoing contribution to Yasuda Kasai's financial performance US\$79 million and US\$9 million in income and total return, respectively, over FY91-92 and it has been in use since then.
3. The basic structure is portable to other applications because of flexible model generation. Indeed, the other models in Table 9.3 are modified versions of the Russell–Yasuda Kasai model.

4. A substantial potential impact in performance of financial services companies.
5. The top 200 insurers worldwide have in excess of $10 trillion in assets.
6. Worldwide pension assets are also about $7.5 trillion, with a $2.5 trillion deficit.
7. The industry is also moving toward more complex products and liabilities and risk-based capital requirements.

I end this section on the stochastic programming approach to asset–liability management by discussing the InnoALM model that I made with Alois Geyer in 2000 for the Siemens Austria pension fund. The model has been in constant use since then and is described in Geyer and Ziemba (2008). I begin by discussing the pension situation in Europe.

There is a rapid ageing of the developed world's populations – the retiree group, those 65 and older, will roughly double from about 20% to about 40% of the worker group, those 15–64.

- Better living conditions, more effective medical systems, a decline in fertility rates, and low immigration into the Western world contribute to this ageing phenomenon.
- By 2030 two workers will have to support each pensioner compared with four now.
- Contribution rates will rise.
- Rules to make pensions less desirable will be made. For example, in the UK, France, Italy, and Greece, they have raised or are considering raising the retirement age.
- Many pension funds worldwide are greatly underfunded.

Historically stocks have greatly outperformed bonds and other assets, see Fig. 9.18 which shows the relative returns in nominal terms. However, the wealth ride is very rocky.

There can be long periods of underperformance. There have been four periods where equities have had essentially zero gains in nominal terms (not counting dividends): 1899–1919, 1929–1954, 1964–1981, and 2000–2008 [seen another way, since 1900 there have only been three periods with nominal gains: 1919–1929, 1954–1964, and 1981–2000].

Among other things, European pension funds, including those in Austria, have greatly underperformed those in the US and the UK and Ireland largely because of they greatly underweight equity. For example, Table 9.7 shows with the asset structure of European pension funds with Austria at 4.1%.

InnoALM is a multiperiod stochastic linear programming model designed by Ziemba and implemented by Geyer with input from Herold and Kontriner for Innovest to use for Austrian pension funds. It is a tool to analyze Tier 2 pension fund investment decisions. It was developed in response to the growing worldwide challenges of ageing populations and increased number of pensioners who put pressure on government services such as health care and Tier 1 national pensions to keep Innovest competitive in their high-level fund management activities.

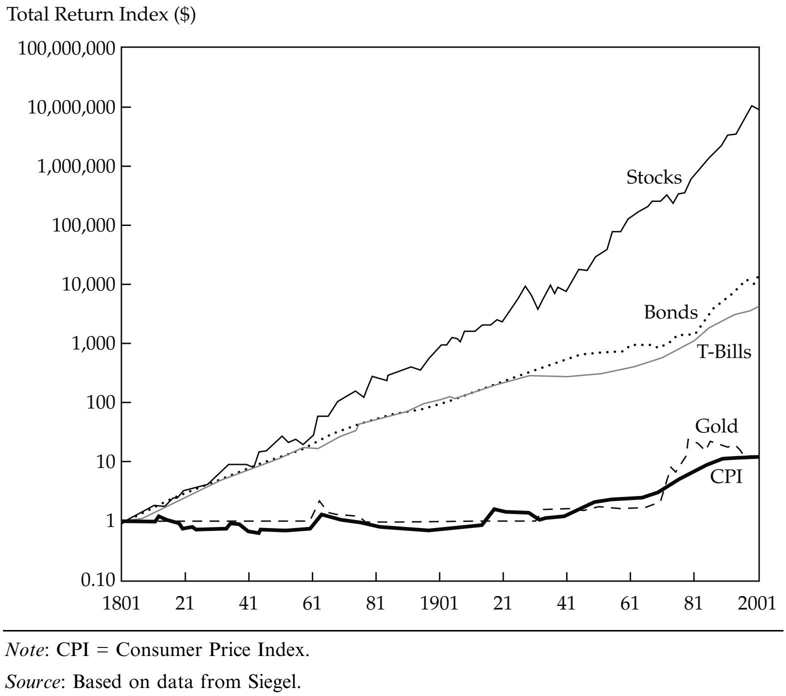


Fig. 9.18 Total nominal return indices, 1802–2001 (Source: Siegel 2002)

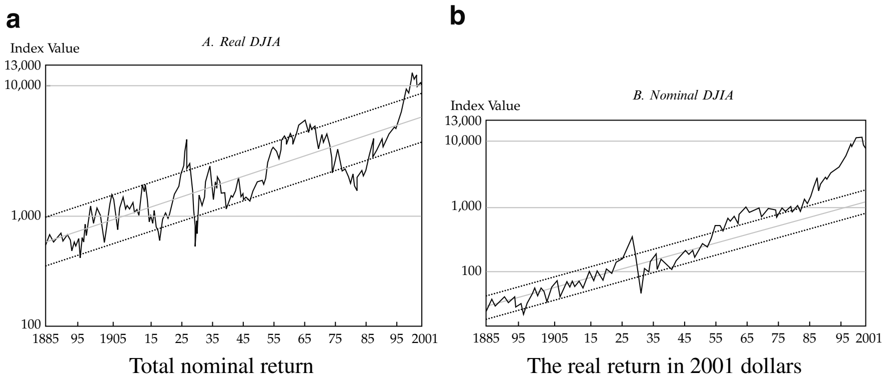


Fig. 9.19 Historical Dow Jones industrial average returns without dividends, 1885 to December 2001 (Source: Siegel 2002)

Siemens AG Österreich is the largest privately owned industrial company in Austria. Turnover (EUR 2.4 Bn. in 1999) is generated in a wide range of business lines including information and communication networks, information and communication products, business services, energy and traveling technology, and medical equipment.

Table 9.7 Asset structure of European pension funds

Countries	Equity	Fixed income	Real estate	Cash STP	Other
Austria	4.1	82.4	1.8	1.6	10.0
Belgium	47.3	41.3	5.2	5.6	0.6
Denmark	23.2	58.6	5.3	1.8	11.1
Finland	13.8	55.0	13.0	18.2	0.0
France	12.6	43.1	7.9	6.5	29.9
Germany	9.0	75.0	13.0	3.0	0.0
Greece	7.0	62.9	8.3	21.8	0.0
Ireland	58.6	27.1	6.0	8.0	0.4
Italy	4.8	76.4	16.7	2.0	0.0
Luxembourg	23.7	59.0	0.0	6.4	11.0
Netherlands	36.8	51.3	5.2	1.5	5.2
Portugal	28.1	55.8	4.6	8.8	2.7
Spain	11.3	60.0	3.7	11.5	13.5
Sweden	40.3	53.5	5.4	0.8	0.1
U.K.	72.9	15.1	5.0	7.0	0.0
Total EU	53.6	32.8	5.8	5.2	2.7
US[a]	52	36	4	8	*n.a.*
Japan[a]	29	63	3	5	*n.a.*

[a]European Federation for Retirement Provision (EFRP) (1996), Table 4

- The Siemens Pension fund, established in 1998, is the largest corporate pension plan in Austria and follows the defined contribution principle.
- More than 15,000 employees and 5,000 pensioners are members of the pension plan with about EUR 500 million in assets under management.
- Innovest Finanzdienstleistungs AG, which was founded in 1998, acts as the investment manager for the Siemens AG Österreich, the Siemens Pension Plan, as well as for other institutional investors in Austria.
- With EUR 2.2 billion in assets under management, Innovest focuses on asset management for institutional money and pension funds.
- The fund was rated the first of 19 pension funds in Austria for the 2-year 1999/2000 period.

Features of InnoALM are as follows:

- A multiperiod stochastic linear programming framework with a flexible number of time periods of varying length
- Generation and aggregation of multiperiod discrete probability scenarios for random return and other parameters
- Various forecasting models
- Scenario-dependent correlations across asset classes
- Multiple covariance matrices corresponding to differing market conditions
- Constraints reflect Austrian pension law and policy

As in the Russell–Yasuda Kasai models, the objective function is a

- Concave risk averse preference function maximizes expected present value of terminal wealth net of expected convex (piecewise linear) penalty costs for wealth and benchmark targets in each decision period.
- InnoALM user interface allows for visualization of key model outputs, the effect of input changes, growing pension benefits from increased deterministic wealth target violations, stochastic benchmark targets, security reserves, policy changes, etc.
- The solution process using the IBM OSL stochastic programming code is fast enough to generate virtually online decisions and results and allows for easy interaction of the user with the model to improve pension fund performance.

The model has deterministic wealth targets that grow 7.5% per year in a typical application. The model also has stochastic benchmark targets on asset returns:

$$\tilde{R}_{\mathrm{B}}B + \tilde{R}_{\mathrm{S}}S + \tilde{R}_{\mathrm{C}}C + \tilde{R}_{\mathrm{RE}}\mathrm{RE} + M_{it} \geqslant \tilde{R}_{\mathrm{BBM}}\mathrm{BBM} + \tilde{R}_{\mathrm{SBM}}\mathrm{SBM} + \tilde{R}_{\mathrm{CBM}}\mathrm{CBM} + \tilde{R}_{\mathrm{REBM}}\mathrm{REBM}.$$

Bonds, stocks, cash, and real estate have stochastic benchmark returns denoted with a ~ asset weights B, S, C, RE with the shortfall Mit to be penalized.

Examples of national investment restrictions on pension plans are as follows:

Country	Investment restrictions
Germany	Max. 30% equities, max. 5% foreign bonds
Austria	Max. 40% equities, max. 45% foreign securities, min. 40% EURO bonds
France	Min. 50% EURO bonds
Portugal	Max. 35% equities
Sweden	Max. 25% equities
UK, US	Prudent man rule

The model gives insight into the wisdom of such rules and portfolios that can be structured around the risks.

- An Excel™ spreadsheet is the user interface.
- The spreadsheet is used to select assets, define the number of periods and the scenario node-structure.
- The user specifies the wealth targets, cash in- and out-flows, and the asset weights that define the benchmark portfolio (if any).
- The input-file contains a sheet with historical data and sheets to specify expected returns, standard deviations, correlation matrices, and steering parameters.
- A typical application with 10,000 scenarios takes about 7–8 min for simulation, generating SMPS files, solving and producing output on a 1.2-GHz Pentium III notebook with 376-MB RAM. For some problems, execution times can be 15–20 min.

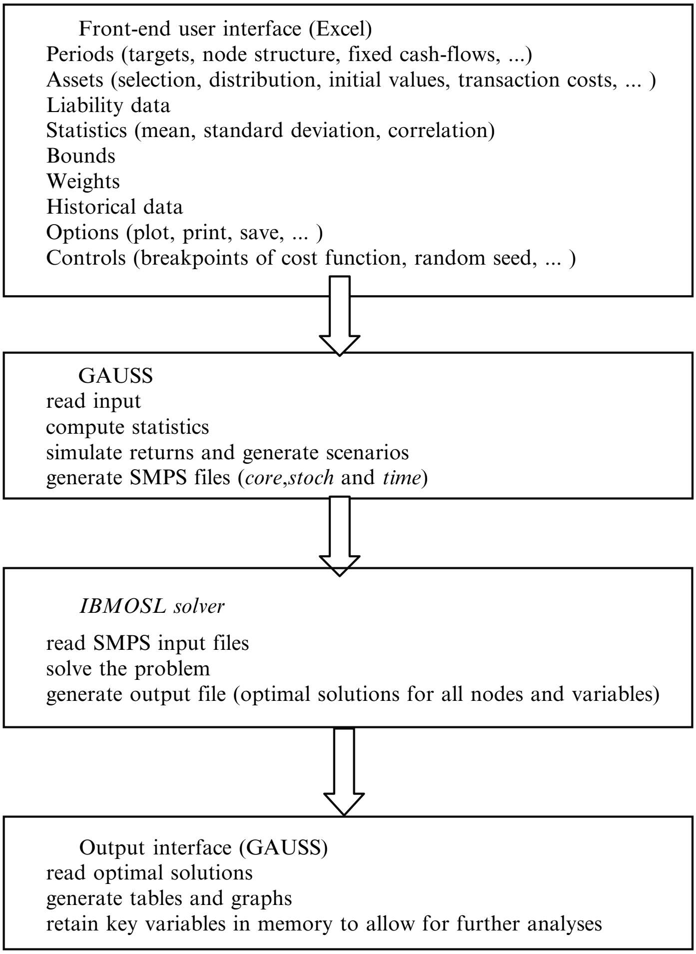
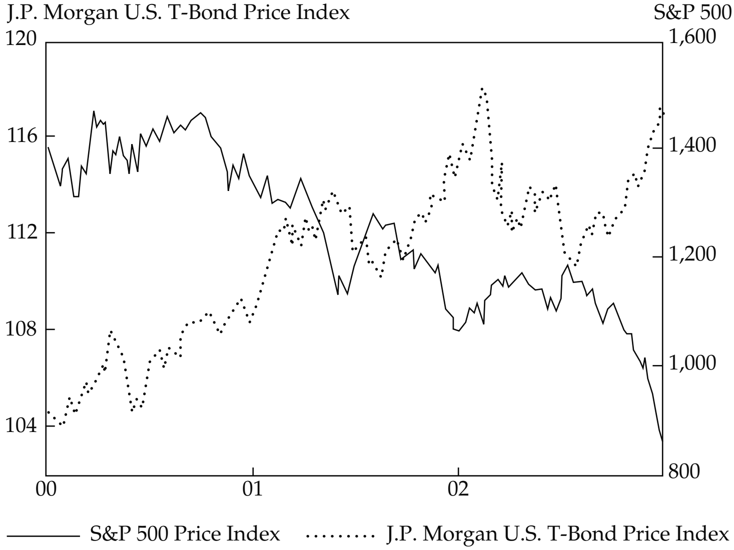


Fig. 9.20 Elements of InnoALM (Source: Geyer et al. 2002)

When there is trouble in the stock market, the positive correlation between stocks and bond fails and they become negatively correlated. When the mean of the stock market is negative, bonds are most attractive as is cash.

Assumptions about the statistical properties of returns measured in nominal Euros are based on a sample of monthly data from January 1970 for stocks and 1986 for bonds to September 2000. Summary statistics for monthly and annual log returns are in Table 9.9. The US and European equity means for the longer period 1970–2000 are much lower than for 1986–2000 and are slightly less volatile. The monthly stock returns are nonnormal and negatively skewed. Monthly stock returns are fat tailed whereas monthly bond returns are close to normal (the critical value of the Jarque–Bera test for $a = 0.01$ is 9.2).

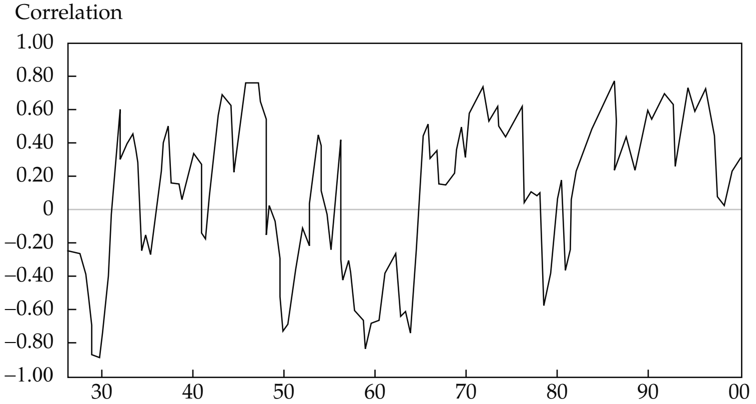


Source: Based on data from Schroder Investment Management Ltd.

Fig. 9.21 S&P500 index and US government bond returns, 2000–2002 (Source: Schroder Investment Management Ltd 2002)

Table 9.8 Means, standard deviations and correlations assumptions based on 1970–2000 data (Source: Geyer and Ziemba 2008)

		Stocks Europe	Stocks US	Bonds Europe	Bonds US
Normal periods (70% of the time)	Stocks US	0.755			
	Bonds Europe	0.334	0.286		
	Bonds US	0.514	0.780	0.333	
	Standard deviation	14.6	17.3	3.3	10.9
High volatility (20% of the time)	Stocks US	0.786			
	Bonds Europe	0.171	0.100		
	Bonds US	0.435	0.715	0.159	
	Standard deviation	19.2	21.1	4.1	12.4
Extreme periods (10% of the time)	Stocks US	0.832			
	Bonds Europe	−0.075	−0.182		
	Bonds US	0.315	0.618	−0.104	
	Standard deviation	21.7	27.1	4.4	12.9
Average period	Stocks US	0.769			
	Bonds Europe	0.261	0.202		
	Bonds US	0.478	0.751	0.255	
	Standard deviation	16.4	19.3	3.6	11.4
All periods	Mean	10.6	10.7	6.5	7.2



Source: Based on data from Schroder Investment Management Ltd.

Fig. 9.22 Rolling correlations between US equity and government bond returns (Source: Schroder Investment Management Ltd 2002)

Table 9.9 Statistical properties of asset returns (Source: Geyer et al. 2002)

	Stocks Eur		Stocks US		Bonds Eur	Bonds US
	1/70 −9/00	1/86 −9/00	1/70 −9/00	1/86 −9/00	1/86 −9/00	1/86 −9/00
Monthly returns						
Mean (% p.a.)	10.6	13.3	10.7	14.8	6.5	7.2
Std. dev (% p.a.)	16.1	17.4	19.0	20.2	3.7	11.3
Skewness	−0.90	−1.43	−0.72	−1.04	−0.50	0.52
Kurtosis	7.05	8.43	5.79	7.09	3.25	3.30
Jarque–Bera test	302.6	277.3	151.9	155.6	7.7	8.5
Annual returns						
Mean (%)	11.1	13.3	11.0	15.2	6.5	6.9
Std. dev (%)	17.2	16.2	20.1	18.4	4.8	12.1
Skewness	−0.53	−0.10	−0.23	−0.28	−0.20	−0.42
Kurtosis	3.23	2.28	2.56	2.45	2.25	2.26
Jarque–Bera test	17.4	3.9	6.2	4.2	5.0	8.7

We calculate optimal portfolios for seven cases. Cases with and without mixing of correlations and consider normal, t-, and historical distributions. Cases NM, HM, and TM use mixing correlations. Case NM assumes normal distributions for all assets. Case HM uses the historical distributions of each asset. Case TM assumes t-distributions with five degrees of freedom for stock returns, whereas bond returns are assumed to have normal distributions. Cases NA, HA, and TA are based on the same distribution assumptions with no mixing of correlations matrices. Instead the correlations and standard deviations used in these cases correspond to an *average' period* where 10%, 20%, and 70% weights are used to compute averages of correlations and standard deviations used in the three different regimes.

Comparisons of the average (A) cases and mixing (M) cases are mainly intended to investigate the effect of mixing correlations. Finally, in the case TMC, we maintain all assumptions of case TM but use Austria's constraints on asset weights. Eurobonds must be at least 40% and equity at most 40%, and these constraints are binding.

A distinct pattern emerges:

- The mixing correlation cases initially assign a much lower weight to European bonds than the average period cases.
- Single-period, mean–variance optimization and the average period cases (NA, HA, and TA) suggest an approximate 45–55 mix between equities and bonds.
- The mixing correlation cases (NM, HM, and TM) imply a 65–35 mix. Investing in US Bonds is not optimal at stage 1 in none of the cases which seems due to the relatively high volatility of US bonds.

Optimal initial asset weights at Stage 1 by case are shown in Table 9.10.

If the level of portfolio wealth exceeds the target, the surplus is allocated to a reserve account and a portion used to increase (10% usually) wealth targets.

Table 9.11 has the expected terminal wealth levels, expected reserves, and probabilities of shortfall. We see that TMC has the poorest results showing that the arbitrary constraints hurt performance. The mixing strategies NM, HM, and, especially, TM have the best results with the highest expected terminal wealthy, expected reserves, and the lowest shortfall probabilities.

In summary, optimal allocations, expected wealth, and shortfall probabilities are mainly affected by considering mixing correlations while the type of distribution

Table 9.10 Optimal initial asset weights at Stage 1 (in %)

	Stocks Europe	Stocks US	Bonds Europe	Bonds US
Single-period, mean–variance optimal weights (average periods)	34.8	9.6	55.6	0.0
Case NA: no mixing (average periods) normal distributions	27.2	10.5	62.3	0.0
Case HA: no mixing (average periods) historical distributions	40.0	4.1	55.9	0.0
Case TA: no mixing (average periods) t-distributions for stocks	44.2	1.1	54.7	0.0
Case NM: mixing correlations normal distributions	47.0	27.6	25.4	0.0
Case HM: mixing correlations historical distributions	37.9	25.2	36.8	0.0
Case TM: mixing correlations t-distributions for stocks	53.4	11.1	35.5	0.0
Case TMC: mixing correlations historical distributions constraints on asset weights	35.1	4.9	60.0	0.0

Table 9.11 Expected terminal wealth, expected reserves, and probabilities of shortfalls with a target wealth, $W_t = 206.1$

Stocks	Stocks Europe	Bonds US	Bonds Europe	US	Expected terminal wealth	Expected reserves at stage 6	Probability of target shortfall	Probability shortfall >10%
NA	34.3	49.6	11.7	4.4	328.9	202.8	11.2	2.7
HA	33.5	48.1	13.6	4.8	328.9	205.2	13.7	3.7
TA	35.5	50.2	11.4	2.9	327.9	202.2	10.9	2.8
NM	38.0	49.7	8.3	4.0	349.8	240.1	9.3	2.2
HM	39.3	46.9	10.1	3.7	349.1	235.2	10.0	2.0
TM	38.1	51.5	7.4	2.9	342.8	226.6	8.3	1.9
TMC	20.4	20.8	46.3	12.4	253.1	86.9	16.1	2.9

chosen has a smaller impact. This distinction is mainly due to the higher proportion allocated to equities if different market conditions are taken into account by mixing correlations.

Geyer and I ran an out-of-sample test to see if the scenario-dependent correlation matrix idea adds value. It was assumed that in the first month, wealth is allocated according to the optimal solution for stage 1. In subsequent months, the portfolio is rebalanced.

- Identify the current volatility regime (extreme, highly volatile, or normal) based on the observed US stock return volatility. It is assumed that the driving variable is US volatility.
- Search the scenario tree to find a node that corresponds to the current volatility regime and has the same or a similar level of wealth.
- The optimal weights from that node determine the rebalancing decision.
- For the no-mixing cases NA, TA, and HA, the information about the current volatility regime cannot be used to identify optimal weights. In those cases, we use the weights from a node with a level of wealth as close as possible to the current level of wealth.

The following quote by Konrad Kontriner (Member of the Board) and Wolfgang Herold (Senior Risk Strategist) of Innovest emphasizes the practical importance of InnoALM:

> The InnoALM model has been in use by Innovest, an Austrian Siemens subsidiary, since its first draft versions in 2000. Meanwhile it has become the only consistently implemented and fully integrated proprietary tool for assessing pension allocation issues within Siemens AG worldwide. Apart from this, consulting projects for various European corporations and pensions funds outside of Siemens have been performed on the basis of the concepts of InnoALM.
>
> The key elements that make InnoALM superior to other consulting models are the flexibility to adopt individual constraints and target functions in combination with the broad and deep array of results, which allows investigation of individual, path dependent behavior of assets and liabilities as well as scenario based and Monte-Carlo like risk assessment of both sides.
>
> In light of recent changes in Austrian pension regulation the latter even gained additional importance, as the rather rigid asset based limits were relaxed for institutions that could

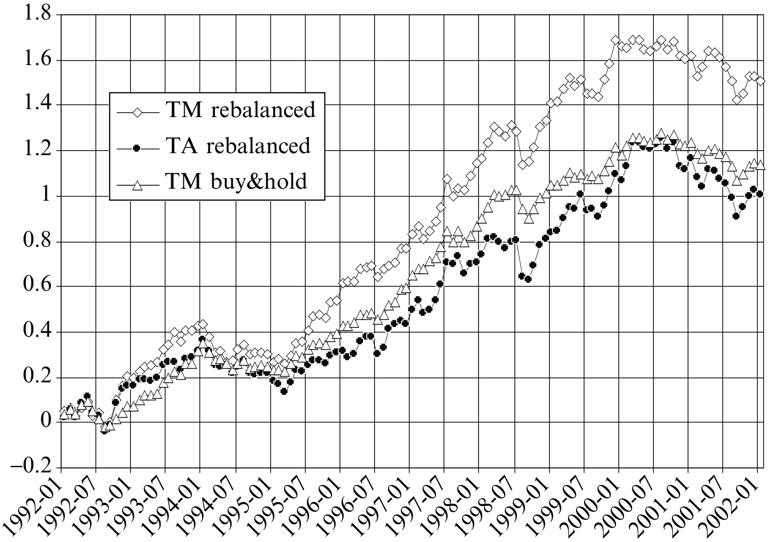


Fig. 9.23 Cumulative monthly returns for different strategies (Source: Geyer and Ziemba 2008)

prove sufficient risk management expertise for both assets and liabilities of the plan. Thus, the implementation of a scenario based asset allocation model will lead to more flexible allocation restraints that will allow for more risk tolerance and will ultimately result in better long term investment performance.

Furthermore, some results of the model have been used by the Austrian regulatory authorities to assess the potential risk stemming from less constrained pension plans.

9.5 Dynamic Portfolio Theory and Practice: The Kelly Capital Growth Approach

The first category of multiperiod models are those reducable to static models. Part IV of Ziemba and Vickson (1975, 2006) reprints key papers and discusses important results.

It is well known in stochastic programming that n-period concave stochastic programs are theoretically equivalent to concave static problems (see Dantzig 1955). In the theory of optimal investment over time, it is not quadratic (the utility function behind the Sharpe ratio) but log that yields the most long-term growth. But the elegant results on the Kelly (1956) criterion, as it is known in the gambling literature and the capital growth theory as it is known in the investments literature, see the surveys by Hakansson and Ziemba (1995) and MacLean and Ziemba (2006), and Thorp (2006), that were proved rigorously by Breiman (1961) and generalized by Algoet and Cover (1988) are long-run asymptotic results. See also the paper of Markowitz (1976), which is reprinted in the full book on the Kelly Capital growth criterion of MacLean, Thorp and Ziemba (2009). However, the Arrow–Pratt absolute risk aversion of the log utility criterion, namely $1/\omega$, is essentially zero. Hence, in the short run, log can be an exceedingly risky utility function with wide swings in wealth values. And fractional Kelly strategies are not much safer as their risk aversion indices are $1/\gamma\omega$ where the negative power utility function is αW^{α}, $\alpha < 0$. This formula

$$\gamma = \frac{1}{1-\alpha}$$

is exact for log normal assets and approximately correct otherwise (see MacLean et al. 2005).

The myopic property of log utility (optimality is secured with today's data and you do not need to consider the past or future) is derived as follows in the simplest Bernoulli trials (p = probability you win and $q = 1 - p$ = probability you lose a fixed amount, where $p > q$).

Final wealth after N trials is

$$X_N = (1+f)^M (1-f)^{N-M} X_0,$$

where f is the fraction bet, X_0 is the initial wealth, and you win M of N trials.

The exponential rate of growth is

$$\begin{aligned}
G &= \lim_{N\to\infty} \log\left(\frac{X_N}{X_0}\right)^{1/N},\\
G &= \lim_{N\to\infty}\left[\frac{M}{N}\log(1+f) + \frac{N-M}{N}\log(1-f)\right]\\
&= p\log(1+f) + q\log(1-f) \text{ by the strong law of large numbers}\\
&= E\log W.
\end{aligned}$$

Thus the criterion of maximizing the long-run exponential rate of asset growth is equivalent to maximizing the one period expected logarithm of wealth. So an optimal policy is myopic. See Hakansson (1971) for generalizations of this myopic policy.

$$\text{Max}\, G(f) = p\log(1+f) + q\log(1-f) \to f^* = p - q.$$

The optimal fraction to bet is the edge $(p-q)$. The bets can be large: $-u''/u' = 1/w \approx 0$

p	0.5	0.51	0.6	0.99
q	0.5	0.49	0.4	0.01
f^*	0	0.02	0.2	0.98

$$f^* = \frac{\text{edge}}{\text{odds}} = \frac{\text{edge}}{10} \quad \text{with 10-1 situation.}$$

The key to the size of the bet is not the edge, it is the risk.

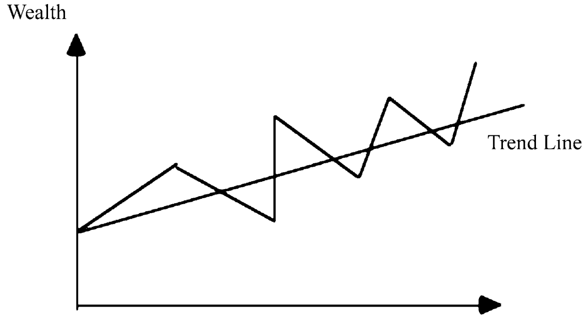


Fig. 9.24 A typical full Kelly wealth path (Source: Ziemba 2003)

An example was the bet to place and show on Slew O'Gold in the inaugural Breeders' Cup Classic in 1984. $f^* = 64\%$ for place/show suggests fractional Kelly (that is a lower bet where you blend the Kelly full-expected log bet with cash). See the discussion of the Thorp–Ziemba bets on that day in Ziemba and Hausch (1986).

In continuous time

$$f^* = \frac{\mu - r}{\sigma^2} - \frac{\text{edge}}{\text{risk(odds)}}$$

$$g^* = \frac{1}{2}\left(\frac{(\mu - r)^2}{\sigma^2}\right) + r$$

$$= \frac{1}{2}\,(\text{Sharpe Ratio})^2 + \text{risk-free asset.}$$

The classic Breiman (1960, 1961) results are: in each period $t = 1, 2, \ldots,$ there are K investment opportunities with returns per unit invested $X_{N_1}, \ldots, X_{N_k}$, intertemporally independent with finitely many distinct values.

$$X_N = \sum_{i=1}^{K} \lambda_i X_{N_i},$$

$$\text{Max } E \log X_N.$$

Property 9.1. Maximizing $E \log X_N$ *asymptotically maximizes the rate of asset growth.*

If in each time period, two portfolio managers have the same family of investment opportunities and one uses a Λ^* maximizing

$$E \log X_N = \sum_{i=1}^{K} \lambda_i X_{N_i},$$

where as the other uses an *essentially different* strategy, e.g.,

$$\text{Max } E \log X_N(\Lambda^*) - E \log X_N(\Lambda) \to \infty$$

then

$$\lim_{N\to\infty} \log \frac{X_N(\Lambda^*)}{X_N(\Lambda)} \to \infty$$

as essentially different means they differ infinitely often. So the actual amount of the fortune exceeds that with any other strategy by more and more as the horizon becomes more distant.

Property 9.2. The expected time to reach a preassigned goal is, asymptotically as X increases, least with a strategy maximizing $E \log X_N$.

We learn a lot from the following Kelly and half Kelly medium time simulations from Ziemba and Hausch (1986). This simulation had 700 independent investments all with a 14% advantage, with 1,000 simulated runs and $w_0 = \$1{,}000$.

Probability of winning	Odds	Expected return	Likelihood of Being chosen in the simulation	f^*
0.57	1 – 1	1.14	0.1	0.14
0.38	2 – 1	1.14	0.3	0.07
0.285	3 – 1	1.14	0.2	0.0475
0.228	4 – 1	1.14	0.2	0.035
0.19	5 – 1	1.14	0.1	0.028

	Final wealth				Number of times the final wealth out of 1,000 trials				
Strategy	Min	Max	Mean	Median	>500	>1000	>10,000	>50,000	>100,000
Kelly	18	483,883	48,135	17,269	916	870	598	302	166
Half Kelly	145	111,770	13,069	8,043	990	954	480	30	1

The power of the Kelly criterion is shown when 16.6% of the time final wealth is 100 times the initial wealth with full Kelly but only once with half Kelly. Thirty percent of the time, final wealth exceeds 50 times initial wealth. But the probability of being ahead is higher with half Kelly 87% vs. 95.4%. So Kelly is much riskier than half Kelly. However, the minimum wealth is 18 and it is only 145 with half Kelly. So with 700 bets all independent with a 14% edge, the result is you can still lose over 98% of your fortune with bad scenarios and half Kelly is not much better with a minimum of 145 or a 85.5% loss. Here financial engineering is important to avoid such bad outcomes.

This is my response to the Samuelson-Merton (1974) and Samuelson (1969, 1971, 2006, 2007) criticism of Kelly betting. See also Luenberger's (1993) response. We agree that you can have a lot of favorable bets and still lose with a Kelly or fractional Kelly strategy. But you can win a lot as well.

9.6 Transactions Costs

The effect of transactions costs which is called slippage in commodity trading is illustrated with the following place/show horseracing formulation (see Hausch et al. 1981). Here q_i is the probability that i wins, and the Harville probability of an ij finish is $q_i q_j/(1-q_i)$, etc. Q, the track payback, is about 0.82 (but is about 0.90 with professional rebates). The players' bets are to place p_j and show s_k for each of the ten horses in the race out of the players' wealth w_0. The bets by the crowd are P_i with $\sum_{i=1}^n P_i = P$ and S_k with $\sum_{k=1}^n S_k = S$. The payoffs are computed so that for place, the first two finishers, say i and j, in either order share the net pool profits once each P_i and p_i bets cost of say \$1 is returned. The show payoffs are computed similarly. The model is

$$\max_{p_i s_i} \sum_{i=1}^{n} \sum_{\substack{j=i \\ j\neq i}}^{n} \sum_{\substack{k=i \\ k\neq i,j}}^{n} \frac{q_i q_j q_k}{(1-q_i)(1-q_i-q_j)} \log \left[\begin{array}{c} \frac{Q(P+\sum_{l=1}^n p_l)-(p_i+p_j+P_{ij})}{2} \\ \times\left[\frac{p_i}{p_i+P_i}+\frac{p_j}{p_j+P_j}\right] \\ +\frac{Q(S+\sum_{l=1}^n s_l)-(s_i+s_j+s_k+S_{ijk})}{3} \\ \times\left[\frac{s_i}{s_i+S_i}+\frac{s_j}{s_j+S_j}+\frac{s_k}{s_k+S_k}\right] \\ +w_0-\sum_{\substack{l=i \\ l\neq i,j,k}}^{n} s_l-\sum_{\substack{l=i \\ l\neq i,j}}^{n} p_l \end{array} \right]$$

$$\text{s.t.} \sum_{l=1}^{n} (p_l + s_l) \leqslant w_0, \quad p_l \geqslant 0, \quad s_l \geqslant 0, \quad l = 1, \ldots, n.$$

While the Harville formulas make sense, the data indicate that they are biased. To correct for this, professional bettors adjust the Harville formulas, using for example discounted Harville formulas, to lower the place and show probabilities for favorites and raise them for the longshots; see papers in Hausch et al. (1994, 2008) and Hausch and Ziemba (2008).

This is a nonconcave program but it seems to converge when nonlinear programming algorithms are used to solve such problems. But a simpler way is via expected value approximation equations using thousands of sample calculations of the NLP model. These are

$$\text{Ex Place}_i = 0.319 + 0.559 \left(\frac{w_i/w}{p_i/p} \right),$$

$$\text{Ex Show}_i = 0.543 + 0.369 \left(\frac{w_i/w}{s_i/s} \right).$$

The expected value (and optimal wager) are functions of only four numbers – the totals to win and place for the horse in question. These equations approximate the full optimized optimal growth model. See Hausch and Ziemba (1985). This is used in calculators.

An example is the 1983 Kentucky Derby.

	Totals	#8 Sunny's Halo	Expected value per dollar bet	Optimal bet (W_0=1,000)
Odds		5-2		
Win	3,143,669	745,524		
Show	1,099,990	179,758	1.14	52

Sunny's Halo won the race

Win	Place	Show
7.00	4.80	4.00

Π = $52

15 second bet!

Watch board in lineup while everyone is at the TV

Here, Sunny's Halo has about 1/6 of the show pool vs. 1/4 of the win pool so the expected value is 1.14, and the optimal Kelly bet is 5.2% of one's wealth.

9.7 Some Great Investors

I end this survey with the wealth plots of a number of great investors who have successfully used the Kelly and fractional Kelly strategies. These include John Maynard Keynes at King's College, University of Cambridge; Warren Buffett of Berkshire Hathaway; Bill Benter the top racetrack bettor in Hong Kong; Ed Thorp, Princeton Newport; and Jim Simons of Renaissance Medallion.

Ziemba (2005) argues that a modification of the Sharpe ratio is needed to evaluate properly the great investors as the ordinary Sharpe ratio

$$S = \frac{\bar{R}_P - R_F}{\sigma_P}$$

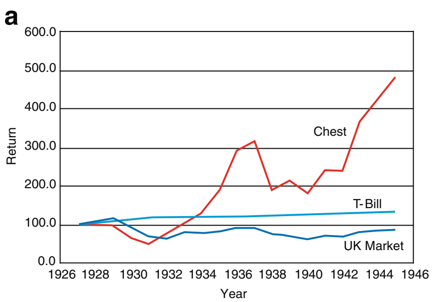
penalizes gains. The modified measure uses only losses in the calculation of σ_P, namely

$$\sigma_P = \frac{\sum_{i=1}^{n}(r_i - \bar{r})^2}{n-1},$$

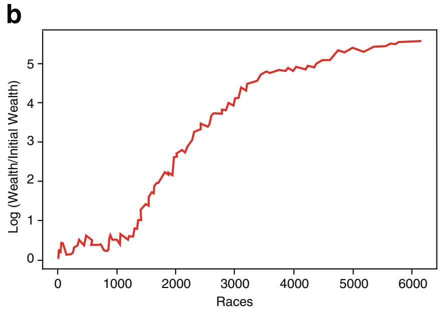
where $\bar{r} = 0$ and ( ) means only the losses are counted. Using this measure, only Buffett improves upon his Sharpe ratio as shown in Table 9.12. But Buffett is still below the Ford Foundation and the Harvard endowment with DSSRs near

Table 9.12 Comparison of ordinary and symmetric downside Sharpe yearly performance measures, monthly data, and arithmetic means (Source: Ziemba 2005)

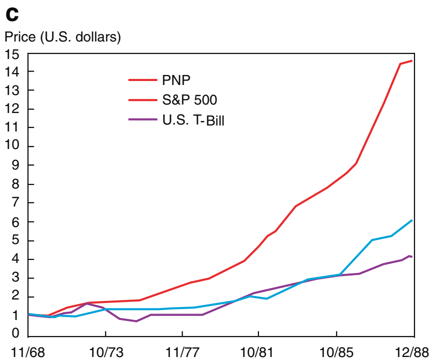
	Ordinary	Downside
Ford foundation	0.970	0.920
Tiger fund	0.879	0.865
S&P500	0.797	0.696
Berkshire hathaway	0.773	0.917
Quantum	0.622	0.458
Windsor	0.543	0.495



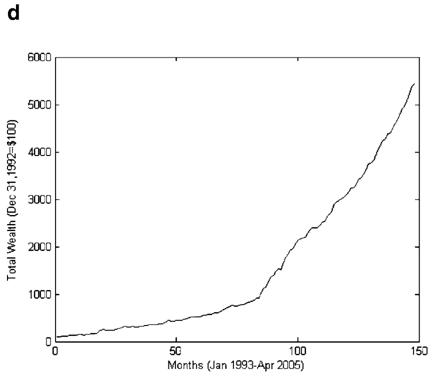
John Maynard Keynes, King's College, Cambridge, 1927-1945 (Source: Ziemba 2003)



Bill Benter, 1989-1991 (Source: Ziemba 2005)



Ed Thorp, Princeton Newport Hedge Fund, 1969-1988 (Source: Ziemba 2003)

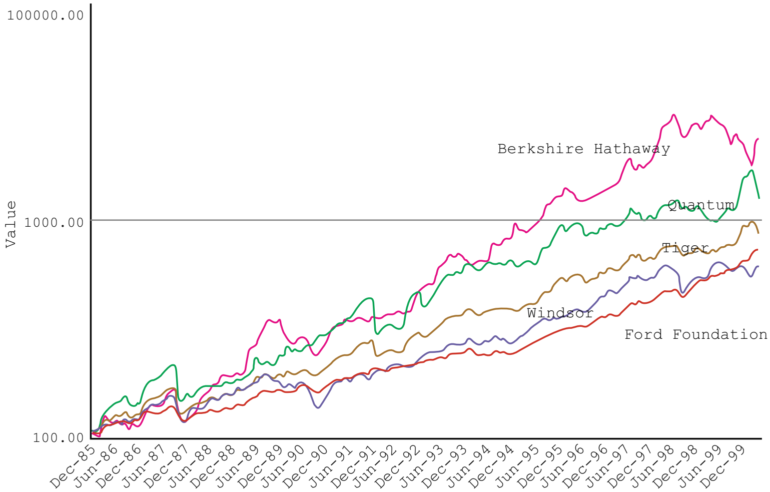


Jim Simons, Renaissance Medallion Hedge Fund, 1993-2005 (Source: Ziemba and Ziemba 2007)

Fig. 9.25 Four great investors

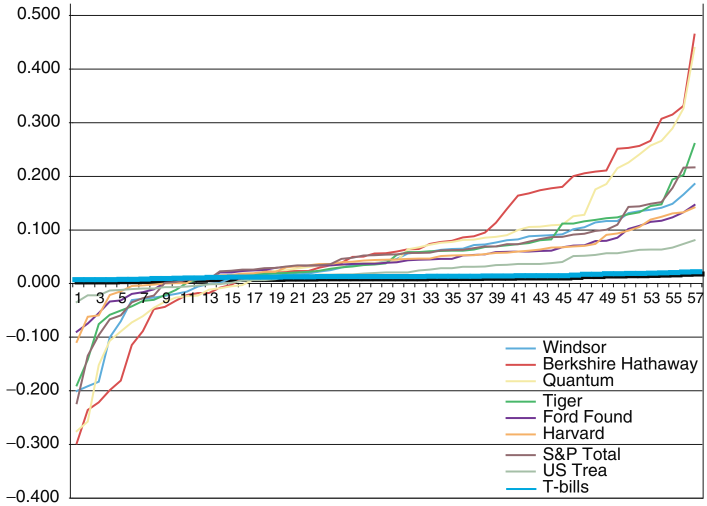
1.00. The reason for that can be seen in Fig. 26(b) where Berkshire has the highest monthly gains but also the largest monthly losses of the funds studies. It is clear that Warren Buffett is a long-term Kelly-type investor who does not care about monthly losses just high final wealth. The great hedge fund investors Thorp at DSSR $= 13.8$ and Simons at 26.4 dominate dramatically. In their cases, the ordinary Sharpe ratio does not show their brilliance. For Simons, his Sharpe was only 1.68.

a



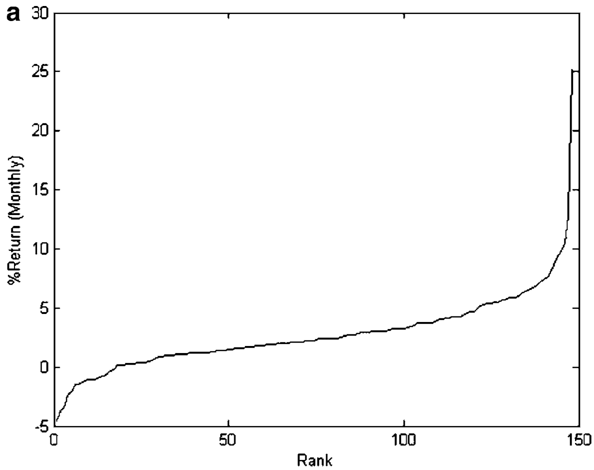
Growth of assets, log scale, various high performing funds, 1985–2000 (Source: Ziemba 2003)

b

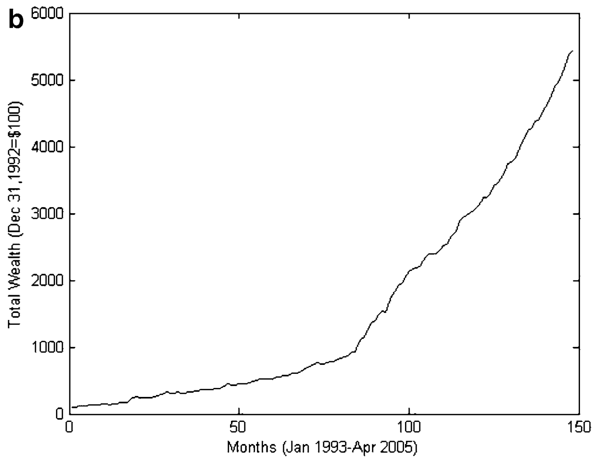


Return distributions of all the funds, quarterly returns distribution, December 1985 to March 2000. Source: Ziemba (2005)

Fig. 9.26 The wealth paths and return distributions of Berkshire Hathaway, Quantum, Tiger, Windsor, the Ford Foundation, and the S&P500, 1985–2000 (Source: Ziemba 2005)



Rates of return in increasing order by month, %



Wealth over time

Fig. 9.27 Medallion fund, January 1993 to April 2005 (Source: Ziemba and Ziemba 2007)

Simons wealth graph, Fig. 27(b) and his distribution of gains and losses, see Fig. 27(a) and Table 9.13 if the best I have seen. His is a remarkable record and its net is of 5% management fees and 44% incentive fees so the gross returns are about double these net returns. Thorp's returns (Fig. 9.25c) are with 2+20 fees and are a model for successful hedge fund performance with no yearly or quarterly losses and only three monthly losses in 20 years.

Table 9.13 Medallion fund net returns, %, January 1993 to April 2005 (Source: Ziemba and Ziemba 2007)

	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
Yearly													
	39.06	70.69	38.33	31.49	21.21	41.50	24.54	98.53	31.12	29.14	25.28	27.77	
Quarterly													
Q1	7.81	14.69	22.06	7.88	3.51	7.30	(0.25)	25.44	12.62	5.90	4.29	9.03	8.30
Q2	25.06	35.48	4.84	1.40	6.60	7.60	6.70	20.51	5.64	7.20	6.59	3.88	
Q3	4.04	11.19	3.62	10.82	8.37	9.69	6.88	8.58	7.60	8.91	8.77	5.71	
Q4	(0.86)	(1.20)	4.31	8.44	1.41	11.73	9.48	20.93	2.42	4.44	3.62	6.72	
Monthly													
January	1.27	4.68	7.4	3.25	1.16	5.02	3.79	10.5	4.67	1.65	2.07	3.76	2.26
February	3.08	5.16	7.54	1.67	2.03	1.96	(2.44)	9.37	2.13	3.03	2.53	1.97	2.86
March	3.28	4.19	5.68	2.77	0.29	0.21	(1.49)	3.8	5.36	1.12	(0.35)	3.05	2.96
April	6.89	2.42	4.10	0.44	1.01	0.61	3.22	9.78	2.97	3.81	1.78	0.86	0.95
May	3.74	5.66	5.53	0.22	4.08	4.56	1.64	7.24	2.44	1.11	3.44	2.61	
June	12.78	25.19	(4.57)	0.73	1.36	2.28	1.71	2.37	0.15	2.13	1.24	0.37	
July	3.15	6.59	(1.28)	4.24	5.45	(1.10)	4.39	5.97	1.00	5.92	1.98	2.20	
August	(0.67)	7.96	5.91	2.97	1.9	4.31	1.22	3.52	3.05	1.68	2.38	2.08	
September	1.54	(3.38)	(0.89)	3.25	0.85	6.33	1.15	(1.02)	3.38	1.13	4.18	1.33	
October	1.88	(2.05)	0.3	6.37	(1.11)	5.33	2.76	6.71	1.89	1.15	0.35	2.39	
November	(1.51)	(0.74)	2.45	5.93	(0.22)	2.26	5.42	8.66	0.17	1.42	1.42	3.03	
December	(1.20)	1.62	1.52	(3.74)	2.77	3.73	1.06	4.30	0.35	1.81	1.81	1.16	

Appendix: Estimating Utility Functions and Risk Aversion

Certainty Equivalent Method

Set $u(w_L) = 0, u(w_U) = 1$

Find $w_{0.5}$

$$u(w_{0.5}) = \tfrac{1}{2}u(w_L) + \tfrac{1}{2}u(w_U)$$

Find $w_{0.25}, w_{0.75}$

$$u(w_{0.25}) = \tfrac{1}{2}u(w_{0.5}) + \tfrac{1}{2}u(w_L)$$
$$u(w_{0.75}) = \tfrac{1}{2}u(w_U) + \tfrac{1}{2}u(w_{0.5})$$

The five points generate u

$$u(w_{0.25}) = 0.25$$
$$u(w_{0.5}) = 0.5$$
$$u(w_{0.75}) = 0.75$$

Split each range one more to generate nine points.

For Donald Hausch, see Fig. 9.2, these values were

w	−30,000	−26,000	−20,000	−14,000	−5,000	+5,000	20,000	28,000	70,000
$u_2(w)$	0.000	0.125	0.250	0.375	0.500	0.625	0.750	0.875	1.000
	↑								↑
	$u(w_L)$								$u(w_U)$

Gain and Loss Equivalent Method

1. Consider the gambles

$$A \begin{cases} \text{lose } \frac{1}{2} \text{ wealth} & \text{pr} = \frac{1}{2} \\ \text{gain} x & \text{pr} = \frac{1}{2} \end{cases}$$

or

$$B \left\{ \text{gain \$5,000 pr=1} \right.$$

. x_0 is the gain that would make the subject indifferent to a choice between A and B (gain equivalence)

$$u(5,000 + w_0) = \frac{1}{2}u(w_0 + x_0) + \frac{1}{2}u\left(\frac{w_0}{2}\right).$$

Set $u(\frac{w_0}{2}) = a, u(w_0 + 5,000) = b.$

Hence this yields $u(w^0 + x_0)$.
2. Repeat with probability of winning = 0.2, 0.3, 0.4, 0.6, and 0.8.
3. Consider the gambles

$$A \begin{cases} \text{lose fraction } y \text{ of wealth} & \text{pr} = \frac{1}{2} \\ \text{gain } x^1 & \text{pr} = \frac{1}{2} \end{cases}$$

or

$$B \left\{ \text{gain \$5,000 pr} = 1 \right.$$

$\rightarrow y_0$ (loss equivalence)

$u(5,000 + w^2) = \frac{1}{2}u(w_0 + x_0) + \frac{1}{2}u(w^0(1 - y^0)).$

Set $u(\frac{w_0}{2}) = a$ and $u(w_0 + 5,000) = b$

4. Repeat with probability of winning = 0.2, 0.3, 0.4, 0.6, and 0.8.
For Donald Hausch, these values were

W	–35,000	–25,000	–15,000	–7,500	+2,500	10,000	14,000	18,000	20,000	39,000	70,000
$u1(w)$	–15.0	–5.00	–3.00	–1.67	0.00	1.00	1.25	1.67	2.00	2.50	5.00
					↑	↑					
					$u(.5\ w_0)$	$u(w_0$+5000)					

Points determined by 1, 2 are above ($w_0 + 5,000$) and by 3, 4 are below (see Fig. 9.2).

Kallberg and Ziemba (1983) Method

Kallberg and Ziemba showed that the average Arrow–Pratt risk aversion approximates closely the optimal Rubinstein measure under the normality assumptions.

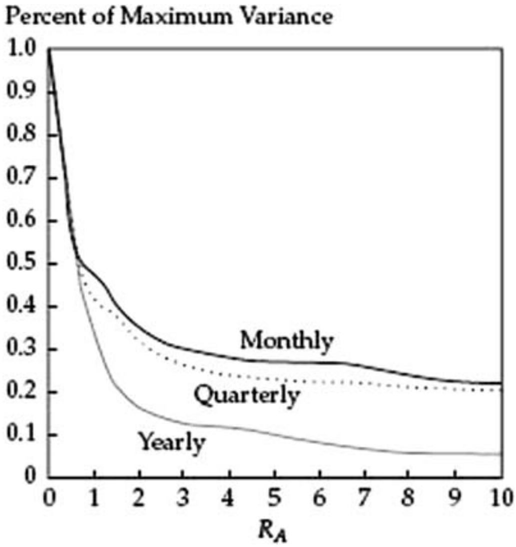


Fig. 9.28 Riskiness as a percentage of maximum variance vs. R_A (Source: Kallberg and Ziemba 1983)

So a consequence is that one can devise risk attitude questions to determine if a particular investor is a $R_A = 4$ (pension fund investor 60–40 mix type) or a risk taker $R_A = 2$ or a conservative investor $R_A = 6$. Kallberg and Ziemba show that the accuracy of such questions is most crucial when R_A is low (see Fig. 9.28).

References

Algoet, P. H. and T. Cover (1988). Asymptotic optimality and asymptotic equipartition properties of log-optimum investment. *Annals of Probability 16*(2), 876–898.

Arrow, K. J. (1965). *Aspects of the Theory of Risk Bearing*. Technical report, Yrjö Jahnsson Foundation, Helsinki.

Breiman, L. (1961). Optimal gambling system for favorable games. *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 63–68.

Chopra, V. K. (1993). Improving optimization. *Journal of Investing 2*(3), 51–59.

Chopra, V. K. and W. T. Ziemba (1993). The effect of errors in mean, variance and co-variance estimates on optimal portfolio choice. *Journal of Portfolio Management 19*, 6–11.

Dantzig, G. (1955). Linear programming under uncertainty. *Management Science 1*, 197–206.

Fleten, S.-E., K. Høyland, and S. Wallace (2002). The performance of stochastic dynamic and fixed mix portfolio models. *European Journal of Operational Research 140*(1), 37–49.

Geyer, A., W. Herold, K. Kontriner, and W. T. Ziemba (2002). *The Innovest Austrian Pension Fund Financial Planning Model InnoALM*. Working paper, UBC.

Geyer, A. and W. T. Ziemba (2008). The innovest Austrian pension fund planning model InnoALM. *Operations Research 56*(4), 797–810.

Gondzio, J. and R. Kouwenberg (2001). High performance computing for asset liability management. *Operations Research 49*, 879–891.

Grinold, R. C. and R. N. Khan (1999). *Active Portfolio Management: Quantitative Theory and Applications*. McGraw-Hill, New York.

Hakansson, N. H. (1971). On optimal myopic portfolio policies with and without serial correlation. *Journal of Business 44*, 324–334.

Hakansson, N. H. and W. T. Ziemba (1995). Capital growth theory. In R. A. Jarrow, V. Maksimovic, and W. T. Ziemba (Eds.), *Finance Handbook*, pp. 123–144. North-Holland, Amsterdam.

Hanoch, G. and H. Levy (1969). The efficiency analysis of choices involving risk. *Review of Economic Studies 36*, 335–346.

Hausch, D. B., V. Lo, and W. T. Ziemba (Eds.) (1994). *Efficiency of Racetrack Betting Markets*. Academic/World Scientific, New York/Singapore; 2nd edn (2008).

Hausch, D. B. and W. T. Ziemba (1985). Transactions costs, entries and extent of inefficiencies in a racetrack betting model. *Management Science XXXI*, 381–394.

Hausch, D. B. and W. T. Ziemba (Eds.) (2008). *Handbook of Sports and Lottery Markets*. Elsevier, Amsterdam.

Hausch, D. B., W. T. Ziemba, and M. E. Rubinstein (1981). Efficiency of the market for racetrack betting. *Management Science XXVII*, 1435–1452.

Kallberg, J. G. and W. T. Ziemba (1981). Remarks on optimal portfolio selection. In G. Bamberg and O. Opitz (Eds.), *Methods of Operations Research, Oelgeschlager*, vol. 44, pp. 507–520. Gunn and Hain, Boston, MA.

Kallberg, J. G. and W. T. Ziemba (1983). Comparison of alternative utility functions in portfolio selection problems. *Management Science 29*(11), 1257–1276.

Kallberg, J. G. and W. T. Ziemba (1984). Mis-specifications in portfolio selection problems. In G. Bamberg and K. Spremann (Eds.), *Risk and Capital*, pp. 74–87. Springer, New York.

Lintner, J. (1965). The valuation of risk assets and the selection of risky investment in stock portfolios and capital budgets. *Review of Economics and Statistics 47*, 13–37.

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, Oxford.

MacLean, L. C., R. Sanegre, Y. Zhao, and W. T. Ziemba (2004). Capital growth with security. *Journal of Economic Dynamics and Control 38*(5), 937–954.

MacLean, L. C., E. O. Thorp, and W. T. Ziemba (2009). The Kelly capital growth investment criterion: Theory and Practice, World Scientific, Singapore.

MacLean, L. C., W. T. Ziemba, and Y. Li (2005). Time to wealth goals in capital accumulation and the optimal trade-off of growth versus securities. *Quantitative Finance 5*(4), 343–357.

MacLean, L. C. and W. T. Ziemba (2006). Capital growth theory and practice. In S. A. Zenios and W. T. Ziemba (Eds.), *Handbook of Asset and Liability Management*, vol. 1, Handbooks in Finance, pp. 139–197. North Holland, Amsterdam.

Mangasarian, O. L. (1969). *Nonlinear Programming*. McGraw-HIll, New York.

Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance 7*(1), 77–91.

Markowitz, H. M. (1976). Investment for the long run: New evidence for an old rule. *Journal of Finance 31*(5), 1273–1286.

Markowitz, H. M. (1987). *Mean–variance analysis in portfolio choice and capital markets*. Basil Blackwell, Cambridge, MA.

Markowitz, H. M. and E. van Dijk (2006). Risk-return analysis. In S. A. Zenios and W. T. Ziemba (Eds.), *Handbook of Asset and Liability Management*, vol. 1, Handbooks in Finance, pp. 139–197. North Holland, Amsterdam.

Merton, R. C. and P. A. Samuelson (1974). Fallacy of the log-normal approximation to optimal portfolio decision-making over many periods. *Journal of Financial Economics 1*, 67–94.

Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica 34*(October), 768–783.

Mossin, J. (1968). Optimal multiperiod portfolio policies. *Journal of Business 41*, 215–229.

Mulvey, J. M. (1996). Generating scenarios for the Towers Perrin investment system. *Interfaces 26*, 1–13.

Mulvey, J. M., B. Pauling, and R. E. Madey (2003). Advantages of multi-period portfolio models. *Journal of Portfolio Management Winter*, *29*, 35–45.

Mulvey, J. M., B. Pauling, S. Britt, and F. Morin (2007). Dynamic financial analysis for multinations insurance companies. In S. A. Zenios and W. T. Ziemba (Eds.), *Handbook of Asset and Liability Management*, vol. 2, Handbooks in Finance, pp. 543–590. North Holland, Amsterdam.

Murty, K. G. (1972). On the number of solutions to the complementarity problem and spanning properties of complementarity cones. *Linear Algebra 5*, 65–108.

Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica 32*, 122–136.

Rockafellar, T. and W. T. Ziemba (2000). Modified risk measures and acceptance sets. Mimeo, University of Washington, July.

Samuelson, P. A. (1969). Lifetime portfolio selection by dynamic stochastic programming. *Review of Economics and Statistics 51*, 239–246.

Samuelson, P. A. (1971). The fallacy of maximizing the geometric mean in long sequences of investing or gambling. *Proceedings of National Academy of Science 68*, 2493–2496.

Samuelson, P. A. (2006). Letter to W. T. Ziemba, December 13.

Samuelson, P. A. (2007). Letter to W. T. Ziemba, May 7.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance 19*, 425–442.

Thorp, E. O. (2006). The Kelly criterion in blackjack, sports betting and the stock market. In S. A. Zenios and W. T. Ziemba (Eds.), *Handbook of Asset and Liability Management*, Handbooks in Finance, pp. 385–428. North Holland, Amsterdam.

Tobin, J. (1958). Liquidity preference as behavior towards risk. *Review of Economic Studies 25*(2), 65–86.

Zangwill, W. I. (1969). *Nonlinear Programming: A Unified Approach.* Prentice Hall, Upper Saddle River, NJ.

Ziemba, R. E. S. and W. T. Ziemba (2007). *Scenarios for Risk Management and Global Investment Strategies*. Wiley, New York.

Ziemba, W. T. (1974). Calculating investment portfolios when the returns have stable distributions. In P. L. Hammer and G. Zoutendijk (Eds.), *Mathematical Programming in Theory and Practice*, pp. 443–482. North Holland, Amsterdam.

Ziemba, W. T. (2003). *The Stochastic Programming Approach to Asset Liability and Wealth Management*. AIMR, Charlottesville, VA.

Ziemba, W. T. (2005). The symmetric downside risk sharpe ratio and the evaluation of great investors and speculators. *Journal of Portfolio Management Fall*, 108–122.

Ziemba, W. T. (2007). The Russell–Yasuda Kasai InnoALM and related models for pensions, insurance companies and high net worth individuals. In S. A. Zenios and W. T. Ziemba (Eds.), *Handbook of Asset and Liability Management*, vol. 2, Handbooks in Finance, pp. 861–962. North Holland, Amsterdam.

Ziemba, W. T. and D. B. Hausch (1986). *Betting at the Racetrack.* Dr Z Investments, Inc., San Luis Obispo, CA.

Ziemba, W. T., C. Parkan, and F. J. Brooks-Hill (1974). Calculation of investment portfolios with risk free borrowing and lending. *Management Science XXI*, 209–222.

Ziemba, W. T. and R. G. Vickson (Eds.) (1975). *Stochastic Optimization Models in Finance*. Academic, New York; Reprinted (2006) With a new preface. World Scientific, Singapore.

Chapter 10
Methodologies for Isolating and Assessing the Portfolio Performance Potential of Stock Return Forecast Models with an Illustration

Bernell K. Stone and John B. Guerard, Jr.

10.1 Introduction and Overview

10.1.1 Forecasting for Mean–Variance Portfolio Optimization

Active portfolio management requires forecasting future returns and/or future risks, generally both. For instance, input to a mean–variance optimizer of Markowitz (1952, 1959) is a vector of predicted asset returns and a matrix of predicted variance–covariance values for the population of assets. Good return forecasting is central to good model-based active portfolio management including especially good implementations of active mean–variance portfolio optimization.

Given the assumption of a competitive near-efficient stock market, much of the return forecasting literature, especially the academic literature, views ongoing successful stock return forecasting as a very difficult if not impossible problem. Timmerman (2008a) is illustrative of this viewpoint in arguing that investors seeking to improve portfolio performance will quickly identify any profitable patterns in past data, trade on this information, and thereby eliminate future ability to profit. Thus, Timmerman argues that the most success that one can expect from a return forecast is transient local forecastability even when one uses sophisticated adaptive forecasting methods.

The Fama–French (1992) three-factor systematic risk model is illustrative of the academic view that realized return dependencies are predominantly a systematic risk effect. The variables size and book-to-market have been added to the CAPM beta as additional systematic risk measures. In contrast to the academic view that the book to market ratio is primarily a systematic risk effect, value managers assert

B.K. Stone (✉)
Marriott School, Brigham Young University, Utah 84602, Provo
e-mail: BernellStone@gmail.com

J.B. Guerard, Jr.
McKinley Capital Management, LLC, 3301 C Street, Suite 500, Anchorage, AK 99503
e-mail: jguerard@mckinleycapital.com

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_10,
© Springer Science+Business Media, LLC 2010

that the book-to-market ratio is among the measures that have predictive value for *alpha performance*, the difference between realized return and a fair return for time and risk.

In commenting on the long period tests of ten alternative time series return forecasts evaluated in Timmerman (2008a), Brown (2008) is not surprised that the forecasts based only on past returns do not perform well, even with the sophistication of adaptively re-estimated structural shifts, Bayesian learning, and neural nets. Brown criticizes absolute accuracy measures as not necessarily relating to portfolio-level portfolio performance. Brown (2008) asserts that testing stock return forecasts should address directly the question of whether the return forecasts can provide economically and statistically significant improvement in portfolio performance. In referring to improvement in portfolio performance, Brown is referring to *alpha performance*.

One response to the Brown criterion is to acknowledge that portfolio performance is the obvious measure of forecast value when the purpose of the stock return forecast is improving active portfolio management. Frequently, the follow-on assertion is that mutual funds, hedge funds, and other actively managed portfolios are already evaluated for performance. It is common to use the Jensen alpha, a Sharpe ratio, and other risk-adjusted performance comparisons between an actively managed portfolio and a pertinent index benchmark. In fact, portfolio performance evaluation is a major business. There is at least implicit evaluation of return forecasting in these risk-adjusted benchmark comparisons. It would not seem hard to adapt these portfolio-benchmark comparisons to assess how a stock return forecast can contribute to portfolio performance.

It is our contention that the conventional pairwise time series performance assessments are low power tests for three reasons:

- High portfolio-benchmark overlap
- Loss of much sample information
- Short assessment period

For instance, assume that an actively managed portfolio seeks to outperform the S&P 500. Assume further that a portfolio manager uses a stock return forecast to tilt away from an exact match of the S&P 500 by replacing 100 of the 500 stocks in S&P 500 but holds 400 of the S&P 500 stocks with these 400 stocks being the same stocks with the same weights as in the S&P 500. The result when one does a comparison with the benchmark is an 80% identical stock return time series since 400 of the 500 stocks in the active portfolio and the S&P 500 are the same. Assume further that forecasts were made for 2000 stocks out of the universe of public companies. From the viewpoint of using sample information, basing a forecast performance assessment on just 100 of 2,000 sample observations means that only 5% of the sample of stock forecasts has gone into the time series performance assessment. By definition of statistical power as the best use of sample information, not using 95% of the forecasted return sample to assess return forecast potential would generally mean a low power test. Our point in this example is that a meaningful test of return forecast performance should be *full sample*. Moreover, in terms of information to noise, assessments should isolate return forecast performance from other stock return performance rather than having a mix of stocks being evaluated for forecast performance value with stocks not being evaluated.

10.1.2 Synthesis: Some Major Forecast Assessment Issues

The following summarize issues pertinent to testing the performance value of stock return forecast:

1. Must separate the impact on the portfolio of a return forecast from any risk forecast and any risk optimization
2. Should move from short-time pairwise portfolio-index or risk-adjusted portfolio performance to higher power sample-level tests that use an entire sample of stock return forecasts rather than implicitly using a small subset of the sample that happens to enter an actively managed portfolio
3. Must isolate well the dependency of realized stock returns on the return forecast where *isolate well* means that there is very little if any impact on realized returns from other return impact variables that are not part of the return forecast including not only systematic risk variables but also any systematic tax effects, and any other return impacts including non-model alpha performance variables
4. Must test not only for significant dependencies of realized return on a well isolated forecast but must also assess realized risk associated with the forecast including the possibility of skewness (asymmetry)
5. Should correct for both outlier and fattail performance and performance distortion effects
6. Should test for a long time period that covers multiple market and business cycles given that many stock return forecasts can be implicit bets on the direction of the market and/or the direction of the economy

Section 10.6 addresses further the Brown (2008) criterion of value for portfolio performance and expands on this already formidable list of must-do, should-do assessment issues to include realistic complications of actual stock return forecasts such as variable parsimony and conditional, component-focused forecasts that focus primarily on relative rank-ordering.

10.1.3 Purposes and Overview

Our primary purposes are to introduce and illustrate three interdependent methodologies for evaluating the potential of a stock return forecast to contribute to portfolio-level alpha performance. These three methodologies are named and described below:

1. *Return-Forecast Isolation*. Given that stock returns depend on a number of systematic risk factors and possibly even systematic differences in the taxation of

dividends and gains, it is clearly necessary to isolate the dependency of realized returns on a stock return forecast from all other variables including especially any dependency on systematic risks and systematic tax effects.

2. *Alpha Performance vs. Return to Risk*. Given that a dependency of realized returns on a return forecast can be a systematic risk in disguise, we need to be able to assess the extent to which a realized return dependency on forecasted return is in fact an alpha performance effect, a systematic risk effect in disguise, or possibly a combination of these two effects. For the case of a combination, we need to be able to resolve alpha performance from systematic risk. Meaningful forecast assessment must make sure that any apparent portfolio performance benefit is not a risk effect in disguise.
3. *Colinearity Distortions from Non-Model Variables*. Since most component-focused stock return forecasts are parsimonious in the sense that they do not use all company-specific return variables, a good forecast performance assessment will isolate the impact of the forecast on returns from not only systematic risk and systematic tax effects but also from non-model alpha performance variables.

We expand on point #3 above by considering the illustrative eight-variable forecast model of Bloch et al. (1993) that is summarized in Exhibit 10.1. The Bloch et al. (1993) forecast model uses eight fundamental ratio variables but does not

Exhibit 10.1 Summary-overview: return forecast model of Bloch, Guerard, Markowitz, Todd, & Xu

We use the eight-variable stock return forecast model of Bloch, Guerard, Markowitz, Todd, and Xu (1993) to illustrate our forecast isolation and assessment methodologies. This exhibit summarizes the eight variables in two categories: ***current*** and ***smoothed***. This exhibit also states a cross sectional regression equation used to assess relative explanatory power of these variables and a return forecast equation. In our implementation of Bloch, Guerard, Markowitz, Todd, and Xu (1993), the weights in the forecast equation are a normalized ten-month significance-adjusted moving average of the prior ten months regression coefficients lagged two months to reflect time delay in reporting annual financial statement data. Section III-D provides a step-by-step summary for estimating the forecast weights.

Fundamental Variables: Current

EP = [earnings per share]/[price per share] = earnings-price ratio
BP = [book value per share]/[price per share] = book-price ratio
CP = [cash flow per share]/[price per share] = cash flow-price ratio
SP = [net sales per share]/[price per share] = sales-price ratio

Fundamental Variables: Current Relative to Average

REP = [current EP ratio]/[average EP ratio over the past five years]
RBP = [current BP ratio]/[average BP ratio over the past five years]
RCP = [current CP ratio]/[average CP ratio over the past five years]
RSP = [current SP ratio]/[average SP ratio over the past five years]

First-Pass Regression Model for a Representative Stock

$$R_s = a_o + a_1 EP_s + a_2 BP_s + a_3 CP_s + a_4 SP_s + a_5 REP_s + a_6 RBP_s + a_7 BCP_s + a_8 RSP_s + \varepsilon_s$$

Formula for Computing the Forecasted Return for Stocks

$$FR_s = w_1 EP_s + w_2 BP_s + w_3 CP_s + w_4 SP_s + w_5 REP_s + w_6 RBP_s + w_7 BCP_s + w_8 RSP_s$$

include any measure of either growth or profitability. To the extent that either growth or profitability can impact alpha performance measures and to the extent that growth or profitability are cross-sectionally correlated with the return forecast of Bloch et al. (1993), meaningful performance assessment must ensure that any apparent alpha performance is not really an artifact of these other non-model variables.

Rather than just summarizing methods and procedures, we seek to make the issues concrete by using as an illustrative multivariate component-focused parsimonious return forecast. To have a meaningful illustration, we need a stock return forecast model for which there is a significant dependency of realized returns on the return forecast for at least some time periods.

For our illustrative forecast model, we have selected the eight-variable forecast model published in Bloch et al. (1993). Exhibit 10.1 defines the eight fundamental ratio variables used in Bloch et al. (1993) both to explain and predict returns. Exhibit 10.1 also summarizes a regression equation used to estimate the relative explanatory value of the variables and a related forecast equation used to predict returns.

Section 10.2 highlights major results. Then Section 10.3 summarizes the Bloch et al. (1993) implementation and testing of the stock return forecasting model, reviews related literature, and provides a step-by-step summary of how we adapt Bloch et al. (1993) to develop a stock return forecast for every sample stock in each of the 456 months of our 1967–2004 study period.

Section 10.4 continues addressing issues in isolating forecast performance value from other variable effects from the perspective of a general conditional return dependency function. We establish that the problem of forecast isolation is an issue of ensuring that the observed return response to forecast relative rank ordering is a *well-defined response subsurface* (well-defined conditional dependency of return on just the return forecast). The idea of a well-defined response subsurface means that we ensure that all of the observations are a conditional dependency on just the forecast variables with all non-model risk, tax, growth, and other non-model return impact variables being controls that are well isolated from having any systematic impact on the cross-section of returns. Implementing forecast isolating well means having a complete set of controls, i.e., a complete set of return impact variables. Exhibit 10.2 lists ten proposed control variables. These controls are defined and discussed in Sect. 10.4.

Section 10.5 develops our key forecast response isolation methodology including the formulation of a *mathematical assignment program* (**MAP**) that transforms (maps) a crosssection of fractile portfolios formed by rank-ordering all stocks in the sample on *return forecast score*, a mapping of the return forecast into the [0,1] interval. The MAP (formulated mathematically in Sect. 10.5) transforms the cross-section of rank-ordered portfolios into an associated cross-section of well-ordered fractile-based control-matched portfolios. The term *control matched* means that

Exhibit 10.2 Summary of Control Variables

Variable Name	Symbol	Variable Definition
Beta	**β**	$\beta = \mathrm{Cov}(R_s - R_o, R_M - R_o)/\mathrm{Var}(R_M - R_o)$ measured over 3 years of past monthly returns, where R_o is the riskless rate and R_M is return on the market index.
Book-to-Price Ratio (Book-to-Market Ratio)	**BP**	Ratio of book value to market value. **Book value** is the latest (at least two-month back) annual CRSP value for ***total common equity***. **Market value** is the current market value of common stock (current stock price times number shares outstanding).
Size (Market Value of Common Stock)	**S**	The market value of common stock at a point in time
Earnings-Price Ratio (Earnings Yield)	**EP**	The ratio of **Net Income** to **market value**, the reciprocal of the price-earnings ratio
Dividend-Price Ratio (Dividend Yield)	**DP**	The ratio of **Annual Dividends** to **Market Value**
Financial Structure	**FL**	The fraction of **Total Investment** provided by debt and preferred stock
Sales Growth	**SAG**	Five-year average sales growth
Sustainable Growth	**SUG**	The three-year average growth of common equity from retained earnings
Return on Investment	**ROI**	The ratio of **Operating Income** (before extraordinary income and expenses) to **Total Investment**
Return on Equity	**ROE**	The ratio of **Net Income** to **Book Value**
Sales Intensity	**SI**	The ratio of **Sales** to **Total Investment**

each portfolio in the cross section has the same portfolio-average value for each of a specified set of control variables. We refer to this associated cross section as a *control-matched cross section*.

Ensuring that each portfolio is matched on the specified controls is accomplished via constraints in the mathematical program. The objective function (developed in Sect. 10.5) is defined to ensure that reassignment of stocks across portfolios optimizes a trade-off between two complementary attributes of statistical power: a set of well-ordered portfolios having a wide range of predicted returns and within-portfolio homogeneity of the return forecast score.

Sections 10.6–10.11 describe our sample, set forth empirical methods and procedures and then present our empirical results. Section 10.12 concludes and notes issues/opportunities for further research.

10.2 Overview of Some Results

In assessing the impact of control variables and the ability to isolate realized return dependency on just the return forecast, we start with a time series of 454 cross-sections of 30 fractile portfolios with no controls. Imposing a control set such as the three risk controls (beta, size, and book to market) means that each of the 454 cross-sections of fractile portfolios is mapped into a corresponding cross-section of control-matched fractile-based portfolios. When there is no variation in portfolio-level values of beta, size, or the book to market ratio, these variables have no differential impact on the cross-section of stock returns. From the viewpoint of having an impact on the cross-section of returns, these control-matched variables are isolated from the return-forecast dependency.

The stepwise imposition of alternative control sets and especially the stepwise imposition of progressively more complete sets of control variables enables an assessment of how the cross-sections of realized returns depend not only on the return forecast but, by observing whether there is a significant change in the return cross section for two difference sets of control variables, we can assess whether the control variables were a source of distortion in the realized return cross section before the control constraints were imposed.

An easy to understand example is to see how the long-run high-minus-low return differences change with the imposition of progressively more complete sets of control variables. In tests developed in Sect. 10.7 and summarized in Exhibit 10.6, the high-minus-low difference for the long-run average annualized return on portfolio 30 (highest return forecast score) less the long-run return for portfolio 1 (lowest return forecast score) is 8.8% with just rank-order grouping with no controls. With the imposition of the three standard risk controls (beta, size, and book-to-market), the long-run average high-minus-low difference is essentially the same, 8.7%. However, imposing just the three tax controls (dividend yield, earnings yield, and fraction of debt financing) without the three risk controls results in a high-minus-low difference of 19.6%. Imposing the three risk and three tax controls together produces a high-minus-low difference of 21.1%. Adding two growth controls and ROE (return on equity, a profitability control) to the six risk-tax controls results in a high-minus-low difference of 26.6%.

For the illustrative stock return forecast of Bloch et al. (1993), the three tax controls are very important for isolating well the return-forecast dependency. Growth-profitability controls are also important. Although the high-minus-low return differences cited in the paragraph above for just the three risk controls had only a modest impact on the long-run high-minus-low difference, the risk controls are important in combination with other controls. Moreover, for the cross sections of realized standard deviations, realized Sharpe ratios, and realized skewness, the three risk controls are important both individually and in combination with other control variables.

The cross-section of realized standard deviations is particularly interesting. The plots of realized standard deviations versus return forecast rank (Exhibits 10.5A through 10.5L) show that standard deviation is a smooth, nonlinear, non-monotonic dependency with greater realized standard deviations in both the low forecast and the high forecast ends of the cross-section. Exhibit 10.5G summarizes the long-run data for the case of the three risk controls and the three tax controls. If we consider how standard deviation varies with return forecast score, then the tabulation (and plot) in Exhibit 5G of realized long-run standard deviations shows values of 6.50, 6.45, and 6.35 for the three portfolios with the lowest return forecast scores. The realized standard deviations then dip below 6%, decline smoothly to a low value of 5.45% for portfolio 14, and then increase slowly to 6-plus percent. Then, for the last three portfolios with the highest return forecast score, the realized standard deviations jump above 7% with values of 7.38, 8.38, and 8.92% for portfolios 28, 29, and 30, respectively. The range in realized standard deviations is less than 3.5%, which is less than one fifth the range in long-run average risky returns. Hence, we expect a wide range in realized Sharpe ratios. For the six risk-tax controls Exhibit 5G, the Sharpe ratios range from essentially zero for the low return forecast scores to 0.2 for the high return forecast score.

Resolving alpha performance from realized risk is complicated by the fact that there is significant realized skewness, negative for low return forecast scores and then turning positive for high return forecast scores. Tests for alpha performance vs. standard deviation risk indicate significant alpha performance (e.g., very significant cross-sectional improvement in Sharpe ratios with return forecast score) and also a significant increase in realized skewness. Most interesting are the stepwise regressions of a realized return on return forecast score and then stepwise additions of quadratic dependencies for both standard deviation and skewness summarized in Exhibit 10.12. When linear and quadratic dependencies of realized standard deviation are added to the return forecast score for well-controlled cross-sections, there is a very significant dependency of realized return on the return forecast score. However, the linear coefficient on standard deviation is significantly negative once both risk and tax controls are imposed. For all the cross-sections that include the six risk-tax controls and especially the cross-sections with risk, tax, growth, and profitability controls, we had an interesting puzzle (discussed further in concluding Section 10.12) of significantly increasing alpha performance with forecast score with a complex nonlinear non-monotonic relatively small but significant cross-sectional variation in realized standard deviation, and an increasing cross-sectional dependence on skewness.

The paragraphs above have discussed the return forecast isolation methodology and the return risk resolution methods in the context of our illustrative forecast. The fact that the cross-section of returns, standard deviations, skewness coefficients, and Sharpe ratios has significant changes with changes in the control sets is indicative of value for the methodologies per se. In the conclusion, we site the high levels of significance as indicative of establishing the merits of the methodologies introduced and tested in the rest of this paper.

Two further comments are pertinent.

1. *Number of control variables*. We find that at least eight control variables are required to isolate well realized stock returns. The need for eight-plus variables to

explain the cross-section of stock returns is considerably more than the conventional 1 to 4 in current systematic risk models of return.

2. *Tax effects*. The finding of significant systematic tax effects is very pertinent for any assessment of cross-sectional return dependency assessments.

10.3 The Illustrative Stock Return Forecasting Model

To illustrate the control-based return forecast isolation methods and the alpha-vs.-risk tests, we need an illustrative forecast model for which there is a statistically significant dependency of realized returns on the return forecast. In the spirit of a volume dedicated to the work of Harry Markowitz, we have selected the eight-variable return forecast framework introduced and tested in Bloch et al. (1993). When used with mean–variance optimization, this model produced significant portfolio performance relative to benchmarks in both Japan and the USA, where "significant portfolio performance" means excess returns that were too great to have been produced by data mining as tested in Markowitz and Xu (1994).

From the viewpoint of illustrating methodology for isolating return forecast performance from other return performance sources, what is required for a good illustrative forecast model is that the cross-section of stock returns exhibits a statistically significant dependency on the return forecast. From the point of view of an illustration, we do not care if the apparent dependency of realized returns on the stock return forecast is alpha performance or is from risk or even if it arises from correlation between the stock return forecast and other non-model return impact variables. In fact, when it comes to illustrating forecast isolation methodology, it is actually good if the crosssectional return dependency is a mixture of effects from the return forecast itself, from systematic risk variables, and from other non-model return performance variables. In effect, to illustrate isolation methodology, it is actually good to have a "dirty return dependency" in the sense that the return dependency includes apparent performance from variables other than the forecast model itself.

Section 10.3.1 defines eight forecast variables. Section 10.3.2 discusses the forecast variables and summarizes major results from Bloch et al. (1993). Section 10.3.3 summarizes the forecast procedure. Sections 10.3.4 and 10.3.5 provide background on value concepts and note the conflict between using value variables to identify misvalued securities for alpha performance and using value as a risk instrument. We review here pertinent multivariate value and value-growth crosssectional return studies. Section 10.3.6 summarizes key results and issues pertinent to isolating well alpha performance from risk, growth, and other return effects.

10.3.1 Notation: Defining Current Value Ratios and Relative Value Ratios

We define four *current value-price ratios* that we refer to as *current value ratios*:

EP = [earnings per share]/[price per share] = earnings-price ratio
BP = [book value per share]/[price per share] = book-price ratio[1]
CP = [cash flow per share]/[price per share] = cash-price ratio
SP = [net sales per share]/[price per share] = sales-price ratio.

We shall generally refer to these ratios as *current value ratios*. The term "current" means the most recent price and the most recent annual report value of earnings, book value, cash flow, and net sales. "Most recent annual report value" means last published annual report lagged at least 2 months after the end of the fiscal year to allow for the delay between the final day of a company's fiscal year and the disclosure of the annual report to the public.

In addition to the current value of each of these four value-price ratios, we define a *relative value ratio* as the current value divided by the most recent 5-year average value of that ratio, i.e.:

REP = EP/[most recent 5-year average value]
RBP = BP/[most recent 5-year average value]
RCP = CP/[most recent 5-year average value]
RSP = SP/[most recent 5-year average value]

10.3.2 Overview of Bloch et al. (1993)

Exhibit 10.1 summarizes the eight company-specific forecast variables. It also states a crosssectional regression equation that can be used to estimate the relative ability of these eight variables to explain the crosssection of realized returns at a point-in-time. Finally, Exhibit 10.1 states a return forecast equation that uses the past regression estimates to construct a relative weighting of the eight value variables. A procedure for determining the relative weights is summarized in detail in Sect. 10.3.3 after discussing the types of value variables and summarizing major results of Bloch et al. (1993).

Exhibit 10.1 contains two types of value ratios – current and relative. These two types of value ratios arise from two complementary ways that value managers say they use value ratios, namely attractiveness relative to peer companies and attractiveness relative to a company's own past valuations.

[1] The ratio BP is of course the book to market ratio. We have defined BP as the ratio of book value per share to price per share. However, multiplying both numerator and denominator by the number of outstanding shares gives the ratio of book value to market value.

Current value ratios measure relative attractiveness compared with other peer companies. All other things being equal, a relatively high EP ratio for a company means the company is relatively more attractive than the peer companies with lower values for their EP ratios. A similar statement can be made for each of the other three current value ratios. Using a combination of four current value ratios together means that information from the set of complementary value ratios enables an assessment of relative attractiveness of one company in comparison with other peer companies.

The four relative value ratios each indicates relative attractiveness compared with a company's own past values. In this case, current value compared with past values is measured relative to the 5-year average of past values. Thus, a stock is viewed as attractive not only when it provides a relatively higher earnings yield than peer companies but also when it provides a high earnings yield relative to its own past values.

To our knowledge, Bloch et al. (1993) are the first to use relative value ratios, either individually or in combination, in a formal stock return forecasting model. More interestingly, except for the extension of their model in Guerard et al. (1997), only Ziemba (1991) tested relative variables in a composite equity forecasting model. Bloch et al. (1993) used the eight-variable forecast model summarized in Exhibit 10.1 to develop rolling horizon predictions of stock returns for both Japan and the USA. They then combined these rolling horizon stock return forecasts with rolling horizon forecasts of risk parameters as input to a mean–variance optimizer [see Markowitz (1959, 1987)] to create mean–variance efficient portfolios in both Japan (first section, non-financial Tokyo Stock Exchange common stocks, January 1975 to December 1990) and the USA (the 1,000 largest market-capitalized common stocks, November 1975 to December 1990).[2] In both Japan and the USA, the mean–variance optimized portfolios significantly outperformed benchmark indices.[3]

For Japan and the USA, Bloch et al. (1993) compared OLS and WLRR (weighted latent root regression) (using both outlier-adjusted and multicollinearity-adjusted regression weighting techniques) by inputting forecasts produced with each method into a mean–variance optimizer. For both Japan and the USA, the portfolio constructed using WLRR-estimated coefficients produced higher Sharpe ratios and geometric means than the portfolio constructed using OLS-estimated coefficients. This result indicates that controlling for both outliers and multicollinearity is important in using regression-estimated composite forecasts.

Bloch et al. (1993) explicitly address two principle issues of forecast value compared with in-sample explanatory value and back-test performance, namely survivor

[2] The use of non-financial stocks led to a customized index for the Markowitz Global Portfolio Research Group. Guerard (1991) presented a research report using this index at the meeting of the Berkeley Program in Finance (Santa Barbara, September 1991). The presentation served as an introduction to the Bloch et al. (1993) model.

[3] Bloch et al. (1993) investigated period-to-period portfolio revision and found that tighter turnover and rebalancing triggers led to higher portfolio returns for value-based strategies.

bias and data mining. They quantified survivor bias and found it was not statistically significant in either Japan or the USA for the 1975–1990 time period tested. Finally, Markowitz and Xu developed a test for data mining [later published in Markowitz and Xu (1994)]. The data mining test allows assessment of the expected difference between the best test model and an average of simulated policies.[4] Guerard and Chettiappan (2009) find statistically significant evidence that their modeling effort and results are not the result of data mining.

10.3.3 The Return Forecast Model

Having identified eight ratio variables as potential return predictors, the forecast modeling question is how to use these variables to predict future returns. In building a composite forecast from a given set of variables, the primary question of composite modeling is how to specify the relative weights.[5] A naïve, all-variables-have-equal-predictive-value approach is to use equal weights, e.g., one-eighth (0.125) for the eight proposed variables.

Rather than the naïve equal-weighting, Bloch et al. (1993) use regression to estimate the relative explanatory power of these variables. Next, they build an adaptively reparameterized stock return forecast model by modifying, smoothing, and normalizing the regression coefficients estimated in recent time periods. The rest of Sect. 10.3.3 summarizes our adaptation of the Bloch et al. (1993) forecast framework to construct the illustrative month-to-month stock return forecast.

We are focusing here on month-to-month returns. With the understanding that we are focusing on the development of a point-in-time return forecast for every stock in a sample, we avoid time subscripts and let the subscript s denote a representative sample stock value. Thus, let R_s denote the total return on stock s. A point-in-time cross-sectional regression equation using these eight variables can be stated as

$$\begin{aligned} R_s = {} & a_0 + a_1\mathrm{EP}_s + a_2\mathrm{BP}_s + a_3\mathrm{CP}_s + a_4\mathrm{SP}_s + a_5\mathrm{REP}_s + a_6\mathrm{RBP}_s + a_7\mathrm{RCP}_s \\ & + a_8\mathrm{RSP}_s + \epsilon_s. \end{aligned} \tag{10.1}$$

[4] Guerard and Chettiappan (2009) report an implementation of this data mining methodology in this volume using an international equity universe.

[5] In referring to the stock return forecast being presented here as a *composite model*, we want to distinguish types of composite forecasting. One type is first to develop two or more separate forecasts and then take a weighted average of the separate predictions. We are not using an averaging of forecasts in the illustrative model of Bloch et al. (1993). Rather, as advocated in Clements and Hendry (1998, especially Chapter 10), we are developing a relative variable weighting. In this case, we use ten cross-sectional regressions delayed 2 months as input to a rolling horizon estimate of relative predictive value over the past 10 months.

Estimating this regression equation in a past time period tells the relative ability to explain past returns.[6] If the return generation structure persists into the future, it could have predictive value. Given regression estimates for a series of past months, additional forecast design questions pertain to forecast horizon and coefficient modification. In particular:

1. If one is going to average (smooth) over a series of past months, how many months should be used?
2. How should estimated coefficients in each month be modified and/or weighted to reflect significance and/or outliers?

The forecast is a rolling average of the robust regression estimates for 10 of the past 12 months. For instance, when we form a forecast score for June 1985, we use regressions from July 1984 through the end of March 1985 but we do not include regression coefficient information from either May 1985 or April 1985 in forming the June 1985 relative variable weighting. In effect, we impose a 2-month delay from the end of the regression estimation period to the start of the return prediction month.

Before taking an average of estimated coefficients over the past 10 months, estimated regression coefficients are modified in three ways in each month:

1. Any coefficient with a t-value ≤ 1.96 is set equal to zero.
2. Extreme positive values are truncated.
3. After steps 1 and 2, the adjusted coefficients are normalized to average to one.

When regression coefficients with t-values ≤ 1.96 are made equal to zero, there are no negative coefficients regardless of significance.

Let $\mathrm{FR_s}$ denote the forecasted return for typical stock s. Let w_i denote the forecast coefficient for the ith value variable. The $\{w_i\}$ are updated monthly, each value being a significance-adjusted, normalized simple average of the past 10 months of regression coefficients. The equation for the forecasted return is for stock s is:

$$\mathrm{FR_s} = w_1\mathrm{EP_s} + w_2\mathrm{BP_s} + w_3\mathrm{CP_s} + w_4\mathrm{SP_s} + w_5\mathrm{REP_s} + w_6\mathrm{RBP_s} + w_7\mathrm{RCP_s} + w_8\mathrm{RSP_s} \tag{10.2}$$

The forecast equation summarized in (10.2) above for developing forecasted returns for each sample stock s is similar to the cross-sectional return regression except that:

1. The regression error term is dropped
2. There is no intercept

[6] Guerard et al. (1997) used a proprietary variable, denoted PRGR, in their study, that combined analysts' 1-year-ahead and 2-year-ahead forecast revisions, forecasted earnings yields, and direction of earnings revisions, denoted as breadth from the work of Wheeler (1991). The PRGR, used as a ninth factor in Guerard et al. (1997) had an average weight of 33% in the composite model for US stocks during the 1982–1994 period. Guerard and Chettiappan (2009) make extensive use of analysts' revisions and find statistical significance in combining fundamental and analysts' expectations data. We discuss this point later in this section.

3. The regression coefficients in (10.1) are replaced by the 10-month average of significance-adjusted, outlier-modified, normalized past coefficient estimates for each variable

For the sake of empirically testing a time series of rank-ordered return cross sections in an extension of the Fama and MacBeth (1973) pooling of CAPM return cross sections, we add a final step not in Bloch et al. (1993). We rank-order the forecasted returns in each month and then scale all returns forecasts by mapping them into the interval (0,1). Or, in percentage terms, we scale all returns forecasts by mapping them into the interval (0%, 100%).

The procedure described above can be summarized in six major steps.

1. Using robust regression, estimate each month the relative ability of the eight variables to explain past return cross-sections.
2. Adjust regression coefficients to reflect significance and/or extreme values and normalize these adjusted coefficients to sum to one.
3. Average adjusted normalized values for each variable from a series of ten past regressions.
4. Update the eight value ratios using variables from the most recent at least 2-month back annual financial statement and current stock prices.
5. Use the coefficients from step 3 and the ratios from step 4 to obtain a return forecast for each stock in accord with equation (10.2) above.
6. Scale the return forecast to map all returns forecasts in step 5 into the interval (0,1) or, in percentage terms, into the interval (0%, 100%).

The forecast score obtained in Step 6 is referred to as a *scaled return forecast score*. Obtaining a scaled return forecast score in Step 6 does nothing to change the relative rank-ordering of predicted performance. Likewise, it does not change entry into fractile portfolios formed by rank-ordering on the basis of predicted return score. The benefit for empirical testing from a normalized return forecast score is that it provides greater cross time comparability of return forecasts. Therefore, the scaled return forecast score provides a return forecast metric other than relative forecast rank for performing cross time empirical tests as discussed further in Sects. 10.7 and 10.8 on empirical testing.

10.3.4 Value-Focused Fundamental Analysis: The Issue of What Value Ratios Measure

Security Analysis, Graham and Dodd (1934), is generally credited with establishing the idea of *value investing*. Although Graham-Dodd analysis involved many qualitative and quantitative considerations (such as a track record of positive and growing earnings and dividends, quality, good asset adequate working capital, correcting for accounting gimmickry while avoiding stocks with excessive debt and poor liquidity), they focused much of their stock-level value assessment on three value attributes: dividends, earnings, and measures of net asset value, all measured relative to the current stock price. The price-earnings ratio tells how much one pays for a dollar of earnings.[7]

Graham and Dodd influenced Williams (1938) who made particular reference to their low P/E and net current approaches in his seminal work, *The Theory of Investment Value*. The reference to Williams (1938) is particularly important because it is *The Theory of Investment Value* that influenced Markowitz's thoughts on return and risk.[8]

Over the past 25 years, value-focused fundamental analysts and portfolio managers have expanded their value measures from primarily price relative to earnings and price relative to book value to also include price relative to cash flow and even price relative to sales. The choice of the eight fundamental variables in Bloch et al. (1993) reflects this expansion in focus, especially the expansion to include cash and sales ratios.[9] Moreover, practice has shifted from comparative use of price multiples to their reciprocals, value relative to price.

The term *value* has multiple meanings. In the context of investment style, "*value*" refers to a *value-growth trade-off* that is often also characterized as *value-glamour trade-off*, as is the work of Lakonishok et al. (1994). In the context of active, performance-focused investing, value (value investing) means selecting (or avoiding) stocks on the basis of comparative measures of value with the implicit assumption being the identification of undervalued or overvalued stocks. Active value investing seeks to select stocks that will produce portfolio performance with

[7] In advocating a low PE strategy, Graham and Dodd (1934) noted that, prior to the 1927–1929 bull market, ten times earnings was a reasonable earnings multiple. By October 1929, many stocks had PE multiples of 25 to 40, a figure that exceeded considerably Graham and Dodd's prudent maximum of 16.

[8] See Markowitz (1991) for an elaboration on this point and Markowitz (1976, 1984, 1987, 1991, and 2000) for further refinements.

[9] The influence of Graham and Dodd, primarily through the low PE multiple, continued in many subsequent works. Students in the 1970s were studying their conclusions in a standard security analysis, investments, and portfolio management text, *Security Analysis and Portfolio Management*, by Latane et al. (1975). The low PE approach is often referred to as an area of "Fundamental Analysis" An investor uses income statement and balance sheet data to estimate an intrinsic value. Such data include sales, earnings, book value, and risk characteristics. Haugen (2001) continues the treatment of the Graham and Dodd in his *Modern Investment Theory*. Haugen (1996,1999) examined 12 of the most important factors in the US equity markets and in Germany, France, Great Britain, and Japan. The book-to-price, earnings-to-price, sales-to-price, and cash flow-to-price variables were among the highest mean payoff variables in the respective countries. Moreover, Graham and Dodd became the standard reference of the value-investing practitioner community, due in a large part to A. Bernhard (1959), the founder of *Value Line*, and his former senior editor, David Dremen, who rose to become chairman of Dremen Value Management, LLC. Dremen (1979, 1998) is the ultimate contrarian. In Dremen (1998), *Contrarian Investment Strategies: The Next Generation*, Dremen showed the vast out-performance of the Low P/E, Low P/CF, Low P/BV (book value), Low P/D (dividends) of the 1500 largest stocks in the USA, 1970–1996 (pp. 154–155). These contrarian variables averaged almost 350 basis points, annually, over the market return; however, Dremen did not account for the riskiness of the strategies.

a positive alpha, i.e., portfolios whose nonsystematic return has a positive expected value. To the extent successful, realized portfolio returns will then have a positive portfolio alpha and a superior Sharpe ratio.

In using value ratios and/or net present values discounted at a risk-adjusted rate, both Graham and Dodd (1934), Williams (1938) and the many value-focused refinements and extensions are seeking to identify stocks with a nonzero expected value for the nonsystematic component of return. Attractive stocks have a positive expected value for the nonsystematic component of return; unattractive stocks have a negative expected value. At the portfolio level, active value investing seeks to select stocks that will produce portfolio performance with a *positive alpha*, i.e., portfolios whose nonsystematic return has a positive expected value and to the extent successful, realized portfolio returns having a positive portfolio alpha.[10]

In contrast to the practitioner focus on the nonsystematic component of return, much of the contemporary academic literature on the cross-section of stock returns uses the term *value* to refer to a systematic risk factor. For instance, in the Fama–French three-factor risk model, *value* refers to one of three systematic risk factors. The book-to-market ratio BP is the variable used to assess value risk. In using the book-to-market ratio BP as a systematic risk measures, the implicit assumption is that there is no risk-corrected performance benefit associated with selecting (avoiding) stocks on the basis of the current value of the book-to-market ratio.

In terms of modeling stock returns, the use of value as a systematic risk factor in the academic literature is clearly a different return concept dealing with a different return component than the use of value ratios to pursue a positive alpha. Whether value is a systematic risk factor or an alpha factor is an empirical question. Moreover, it is very likely that reality is not either just a systematic risk factor or just an alpha factor but a combination of both.

10.3.5 Past Research Pertinent to Regression-Estimated Combinations of Value Ratios

In addition to the extensive literature on the dividend yield tilt that we review in Sect. 10.4, there have been literally hundreds of studies of how the cross-section of stock returns depends on various individual ratios, especially the EP and BP ratios, often in conjunction with dependencies on size (market capitalization) and/or

[10] Moreover, Haugen and Baker (2009) employ many of these variables. There is strong support for both fundamental variables (earnings, book value, cash flow, sales, dividends) and earnings expectations variables in the anomalies literature [Levy (1999), Brown (1993, 2008), and Ramnath et al. (2008)].

growth. Early research includes Basu (1974, 1977, 1983), Banz (1981), Banz and Breen (1986), Jaffe et al. (1989), Chan et al. (1991), Lakonishok et al. (1995), and Fama and French (1992, 1995).[11]

Here we restrict our review to a very specialized subset of value-focused stock return forecasting, namely one-step[12] regression-estimated composites of three or more value ratios, especially those that go beyond using just one or two of the standard value ratios (DP, EP, and/or BP) to include as well the cash-price ratio (CP) and/or the sales-price ratio (SP). In contrast to the extensive academic literature on the impact on individual value ratios on the cross-section of stock returns, the academic literature in this area is relatively limited. In addition to Bloch et al. (1993), the major papers using multiple value ratios to predict/explain a time series of stock return cross sections that include at least CP and/or SP are Chan et al. (1991), Lakonishok et al. (1994), and Guerard et al. (1997). We review these papers in some detail both to summarize key results and to identify unresolved forecast performance assessment issues.

Guerard et al. (1997) extend the fundamental forecasting model in Bloch et al. (1993) in two ways:

1. Adding to the eight-variable regression equation a growth measure.
2. Adding measures of analysts' forecasts and forecast revisions from the I/B/E/S data base, namely consensus analysts' forecasts, forecast revisions, and a measure of the direction (net up or down) of the forecast revisions.

In quarterly weighted latent root regressions applied to the intersection of Compustat, CRSP, and I/B/E/S data bases, the growth variable averaged a relative weight of 33% whereas the average relative weighting of the eight value variables averaged almost 67%. This result complements that of Lakonishok et al. (1994) reviewed below in showing that rank-ordered portfolio returns have both a significant value and growth component.

Adding I/B/E/S variables to the eight value ratios produced more than 2.5% of additional annualized average return. The finding of significant predictive performance value for the I/B/E/S variables indicates that analyst forecast information

---

[11] For more recent work, see Fama and French (2008a, b).

[12] Rather than the one-step direct forecast of stock returns considered here, most of the literature (including the framework presented in investment management texts) and a majority of the stock return forecasting using valuation multiples is in the context of a two-step return forecast in which an analyst predicts both a future value of a variable such as earnings and an associated future value multiple for that variable such as a future price-earnings ratio. The most common version is a prediction of future earnings and a prediction of a future earnings multiple. These two predictions imply a predicted future value. Under the assumption that the current price will converge toward this predicted future value, there is an implied prediction of a gain return. Given a prediction of future dividends, there is an implied stock return forecast. Rather than the conventional two-step procedure, this research uses a direct one-step prediction of return as a combination of current and relative value ratios. For a thorough treatment of the two-step framework and extensive references to the two-step return prediction literature, readers are referred to the CFA study guide by Stowe et al. (2002, 2007).

has value beyond purely statistical extrapolation of past value and growth measures. Possible reasons for the additional performance benefit could be that analysts' forecasts and forecast revisions reflect information in other return-pertinent variables, discontinuities from past data, and/or serve as a quality screen on otherwise out-of-favor stocks. The quality screen idea is a confirmation of the Graham and Dodd argument that value ratios should be used in the context of the many qualitative and quantitative factors that they argue are essential to informed investing.

To test the risk-corrected performance value of the forecasts, Guerard et al. (1997) formed quarterly portfolios with risk being modeled via a four-factor APT–based model (created using 5 years of past monthly data). The portfolios quarterly return averaged 6.18% before correcting for risks and transaction costs with excess returns of 3.6% after correcting for risk and 2.6% quarterly after subtracting 100 basis points to reflect an estimate of two-way transactions costs. The annualized after-risk, after-cost return of 10.4% is both economically and statistically significant. However, while there may have been *implicit* tax corrections in the APT factors, there was no explicit correction for differential taxation of dividends and gains.

Chan et al. (1991) do not formulate a forecast model as such but do perform tests indicating relative explanatory value if not predictive potential. They used seemingly unrelated regression to model CAPM monthly excess returns as functions of EP, BP, CP and size. They estimated an adaptation of the Fama–MacBeth (1973) time series of portfolio cross sections for the Japanese market. These tests produced negative and statistically significant coefficients on EP and size but positive and statistically significant coefficients for the BP and CP variables. Chan et al. (1991, p.1760) summarize their findings with: "The performance of the book to market ratio is especially noteworthy; this variable is the most important of the four variables investigated." These early tests show clear explanatory value to value ratios, especially BP and CP. However, given their CAPM focus, it is not clear whether the significance is from non-beta systematic risk, possibly from tax affects, or from growth effects as investigated in Lakonishok et al. (1994).

In a thorough assessment of value vs. growth in Japan and the USA, Lakonishok et al. (1994) examine the intersection of Compustat and CRSP databases for annual portfolios for NYSE and American Stock Exchange common stocks, April 1963 to April 1990. Their value measures are three current value ratios: EP, BP and CP. Their growth measure is 5-year average annual sales growth (GS). They perform three types of tests: a univariate ranking into annual decile portfolios for each of the four variables, a bivariate ranking on CP (value) and GS (growth, glamour), and finally a multivariate regression adaptation of the Fama and MacBeth (1973) time series pooling of crosssectional regressions. The average high-minus-low spreads for the univariate decile rankings were 10.5, 11.0, 7.6, −2.4% for BP, CP, EP and GS, respectively. In the two-way CP-GS bivariate sort, the extreme value portfolio (highest CP, lowest GS) returned 22.1%, whereas the glamour portfolio (highest GS, lowest CP) returned only 11.4%, a difference of 10.7%. Moreover, low CP stocks with low historic sales growth produced lower portfolio returns than low CP stocks with high past sales growth by over 4.5%.

Lakonishok et al. (1994) used the Fama–Macbeth methodology to construct portfolios and pool (average over time) a time series of 22 1-year cross-sectional univariate regressions for each of the 22 years in their study period. The univariate regression coefficient for SG was significantly negative. The EP, BP, and CP coefficients were all significantly positive. When they performed a multivariate regression using all four variables, they report significantly positive coefficients for BP and EP (but not CP) and significantly negative for SG. Overall Lakonishok et al. (1994) conclude:

1. Buying out-of-flavor value stocks outperformed growth (glamour) over the April 1968 to April 1990 period.
2. Future growth was difficult to predict from past growth alone.
3. The actual future growth of the glamour stocks was much lower than past growth relative to the growth of value stocks.

They test for realized risk and report that the value strategies ex post were not significantly riskier than growth (glamour) strategies ex post. This result raises a question as to whether value is an ex post risk effect or whether growth also should be viewed as a risk effect.

10.3.6 Using Multiple Fundamental Variables in an Adaptively Updated Rolling Horizon Forecast Model: Synthesis

Value ratios individually and together have value in both explaining and predicting the cross-section of stock returns. A combination (composite) of several value ratios seems better than individual ratios. While only addressed in two studies, adding relative ratios (current value relative to own past value) seems to have both explanatory and predictive power. Growth appears to be difficult to predict but clearly interacts with value ratios.

The central question is how much of the performance benefit is risk-based? In using BP as a risk instrument in the Fama–French three-factor model, the implicit assumption is that the BP return impact is a risk reward. This assumption is plausible for any easy to obtain and use variable that produces longrun return benefits but should be tested thoroughly.

Contrary to the implicit Fama–French assumption that BP is all risk, the studies of composites of value ratios each suggest some performance value after correcting for risk. Despite the positive evidence, there is still a need to test thoroughly the question: Can a forecast based on a composite of value ratios add active performance value after correcting for risk and other known return performance factors such as taxes (the well-known dividend yield tilt) and future growth?

10.4 Isolating Forecast Performance Benefits from Systematic Risks, Tax Effects, Growth, and Other Variables

10.4.1 Assessment Complexities for Conditional Component-Focused Forecasts

In Sect. 10.1, we cited the Brown (2008) criterion of improvement in portfolio performance as a much more pertinent measure of forecast value than the standard absolute accuracy measures.

We extended concerns with absolute accuracy measures to: (1) "active alpha" forecasting where the forecasting concern is, by definition, almost always the non-systematic component of stock return, (2) any stock return forecasting situation in which the purpose is a conditional relative rank-ordering, and (3) parsimonious models that intentionally forecast only a subset of the known return dependency variables. The three points below expand further on these limitations in the traditional absolute accuracy tests of return performance by considering three return forecast complexities pertinent to most active performance-focused stock return forecasts including the return forecasting model that we use in this paper to illustrate the performance assessment methods that set forth here.

1. *Relative Rank-Ordering*. In many forecast situations, return forecasts are not concerned with absolute forecast accuracy at all but rather seek to rank-order relative return performance across a population of stocks having a time-varying mean return.
2. *Component-Focused vs. Total-Return-Focused Forecasting*. In many forecasting situations, including the stock return forecasts tested here, the concern is not only relative rank-ordering but also forecasting primarily (if not exclusively) the non-systematic component of return under the assumption of no competitive advantage in predicting the market return (in a CAPM return framework) or no competitive advantage in predicting the systematic return factors in a multifactor return model. Therefore, until one controls for or otherwise isolates the nonsystematic return performance from the systematic return performance, one cannot realistically assess either forecast performance itself or, more importantly, the ability of the component-focused relative rank-ordering to contribute to portfolio performance. The need to isolate nonsystematic return forecast performance from the systematic component is especially important but difficult because the systematic return movements are primarily risk and are generally much larger than the unsystematic return values.
3. *Variable Parsimony*. The eight variable model summarized in Exhibit 10.1 illustrates variable parsimony in the sense that it uses a small subset of the many company-specific return impact variables that could be used.[13] For instance, this

[13] While there is an extensive literature on parsimony in macro forecasting, we are not aware of an analogous literature in the context of stock return forecasting for active portfolio management. Thus, we refer readers to the excellent treatment in Clements and Hendry (1998, Chap. 12, "Parsimony").

forecast model does not include the fundamental value ratio dividend yield (for reasons explained later), does not use growth, and does not use either beta or firm size, two well-established systematic risk variables. To the extent that explanatory variables in the model of Exhibit 10.1 may be correlated with other return impact variables, it is necessary to isolate their impact on return performance from all other return impact variables, especially those non-model variables that are relatively highly correlated with the return forecast score. Beta, size, dividend yields, and growth are non-model return impact variables that are in fact cross-sectionally correlated in most months with the return forecast predictions generated by the model in Exhibit 10.1.

10.4.2 Implementing the Brown Criterion of Portfolio Performance

Moving away from absolute accuracy is crucial for the value models tested here. As we have already noted, any alpha-type performance value is in the unsystematic component of return. In terms of benefiting from any information about the unsystematic component of return, the systematic market movements are noise from which the forecast information must be isolated. Rather than predicting the absolute value of future returns, the fundamental variable models tested here focus on relative ranking without concern for predicting the future value (or even necessarily future direction) of the market itself.

There are two complementary ways to implement the Brown criterion of improved portfolio performance. One is to see whether a return forecast enables portfolio selection/revision that produces a superior return for time and risk. The other is to see whether the use of a return forecast enables the creation of a portfolio that consistently and significantly outperforms a benchmark matching (index matching) no-forecast strategy.

As Timmerman (2008b) notes in his response to the comments on his paper, assessing forecast impact on portfolio performance is complex whether one compares performance to an index benchmark or to a risk-adjusted return. Any portfolio performance assessment is a joint test of the return forecast model, the portfolio selection/revision model used to convert forecasts into portfolios, and an asset pricing framework that defines a fair return for time and risk. Even when comparing portfolio forecast performance to an index benchmark, care must be taken to ensure that the effect of the forecast is not just to select higher risk stocks.

Although the fact that a portfolio performance assessment of a return forecast model is a joint test of the forecast model and an asset pricing framework is generally well understood, it is less obvious that there is also an implicit test of a portfolio selection/revision procedure. To illustrate, Bloch et al. (1993) formulate a return forecast model that projects the relative ability of the eight value ratios listed in Exhibit 10.1 to explain returns. The predicted returns are input along with predictions of return variances and covariances into a mean–variance portfolio optimizer (see Markowitz 1959). Even using just past variance and covariance values (which are reasonably stable relative to period-to-period returns), producing a time series of mean–variance optimized portfolios with Sharpe ratios that are superior to the Sharpe ratios for a benchmark index may be from the return forecast identifying superior risk-adjusted returns, or the superior Sharpe ratios may arise from a lower standard deviation arising from the use of variance–covariance information, or possibly from both. Any conclusions about the ability of a return forecast model to produce superior risk-adjusted performance must be clearly separated from the performance impact of the risk forecast whenever one bases portfolio selection/revision on both a return forecast and a risk forecast including even an implicit risk forecast based on past variance–covariance values.

The point of this discussion is to note the complexity of isolating the contribution of a return forecast from other sources of imputed portfolio performance. When it comes to empirical testing of a period-to-period (month-to-month) forecast model across time, this complexity is magnified considerably when we recognize that we have a conditional asset pricing model in each month with both conditional risks and conditional risk premia in each month. The complexity of time-varying risks and risk premia and other concerns in isolating forecast value from other cross-time return impacts is addressed further after we formulate a general conditional return model and recognize that implementing well the Brown performance criterion is isomorphic to the problem of using observational data to obtain a well-isolated forecast subsurface from a time-varying multivariate and conditional return dependency function.

10.4.3 A General Functional Representation of the Cross-Section of Risky Returns

Let $\{V1,\ V2\ldots\ VJ\}$ be a set of J stock-specific variables that impact stock return. Examples of stock-specific variables include beta, dividend yield, size, growth, financial structure, and of course all eight of the fundamental ratios used in Bloch et al. (1993). From the list of eight model variables in Exhibit 10.1 and the ten control variables summarized in Exhibit 10.2, it is clear that J is a relatively large number.

Let $\{Q1,\ Q2,\ldots,QK\}$ be a set of K market-wide variables. These include not only possible macro conditioning variables but also especially variables such as the market risk premium for pricing beta risk, small minus big for pricing size, and the market tax penalty for return received as dividends rather than gains.

If R_{s} denotes the return on stock s and R_{o} denotes a fair return for time, then risky return on stock s can be represented as

$$R_{\mathrm{s}} - R_{\mathrm{o}} = g(V1, V2\cdots VJ|Q1, Q2\cdots QK) \tag{10.3}$$

The point-in-time CAPM is a very special case in which $J = K = 1$. The only firm-specific variable is a possibly time-varying beta; the one market-wide variable is its price, a time-varying equity risk premium. The Fama–French model involves three firm-specific variables: beta, size, and the book-to-market ratio and three market-wide prices for each of these variables. For the Fama–French three-factor model, we have, of course, $J = K = 3$.

10.4.4 Return Dependency as a Response Surface/Subsurface

We use the term *return response surface*[14] to refer to the functional dependency of return on explanatory variables. In (10.3), we are focusing on a $J + K$ variable risky return response surface. In response surface terms, the function $g(V1,\ V2,\ldots,\ VJ|\ Q1,\ Q2,\ldots,\ QK)$ is a J-variable conditional response surface, i.e., a dependency showing how risky return varies cross-sectionally with variation in the J firm-specific variables $\{V1,\ V2,\ldots,VJ\}$ conditional[15] on $\{Q1,\ldots,QK\}$.

To isolate return forecast performance from other return impact variables, we need to assess how risky return depends on just the forecast. In effect, we need to estimate a cross time conditional dependency. In response surface terms, we need to estimate empirically a risky return response subsurface telling how risky return depends on the forecast alone with all other return impact variables isolated well from the forecast impact.

Our proposed methodology for isolation is to assess how realized risky returns depend on the return forecast for a time series of 30 forecast-rank-ordered fractile portfolios with all other pertinent non-forecast variables set at their population (sample) average values. In effect, all portfolios in each cross section in the time series are matched to the sample average values for control variables. With all control variables set at their sample average values, the portfolio-to-portfolio differences in realized return should be from differences in the return forecast values, random error, or variables omitted from the control set. With a complete set of controls, each

---

[14] *Response surface methods* refer to empirical methods for estimating functional dependencies and especially conditional dependencies (*response subsurfaces*). Most of the statistical literature pertains to controlled or semi-controlled experiments in process industries, medicine, and advertising. For background readers are referred to the foundation works by Box (1954) and Box and Draper (1959), to the review article by Myers et al. (1989), and to the books on response surface methods by Myers (1978), Khuri and Cornell (1996), and Box and Draper (1987, 2005). In Stone et al. (1993), the use of response surface methods is extended to observational data for those estimation situations in which it is pertinent to group data, e.g., to group observations on individual stocks into portfolios. The use of control variables to assess a conditional dependency (response subsurface) is a basic technique in controlled experiments that we adapt in this study to assess portfolio-level conditional return dependencies.

[15] There is recognition that the price of systematic risk factors in the traditional systematic factor-based asset pricing models should be viewed as **conditional** in the sense of time-varying factor prices and/or as conditional in the sense of risk factors depending on additional macroeconomic or firm-specific variables. See Rosenberg (1974), Rosenberg and Rudd (1982), and Brown and Weinstein (1983) for early works in this framework. For more recent explicit conditional frameworks, see for instance Adrian and Rosenberg (2008) and Ferson and Harvey (1991a, 1991b, 1995, 1998).

cross section should be a well defined return-forecast response subsurface. Thus, for a forecast focused on relative rank-ordering part of the non-systematic component of return, we should be able to assess how effectively the forecast produces well-ordered differences in realized risky returns for the entire cross-section of fractile portfolios.

10.4.5 Cross-Sectional Return Controls

We want to be sure to have a complete set of return controls. We have already argued that the number of necessary controls is greater than just the usual systematic risk variables when the return model includes return impact factors beyond non-diversifiable risk such as tax effects, company valuation variables such as financial structure, and possible return performance explainers/predictors such as value ratios and growth measures.

10.4.6 Risk Controls

The CAPM *beta* is measured here using 3 years of monthly risky returns, where risky return is realized return in a month less the monthly T-bill rate for that month. *Size* is simply stock market value (price per share multiplied by total shares outstanding). The current book-to-market ratio *BP* is measured as the ratio of the book value in the most recent (2 month or older) annual report to end-of-month stock market value. It is a variable in the return forecast but it also used here as a risk control. Its contribution to realized return can be set at the same value for every portfolio in the cross section so that we can assess relative portfolio performance with the impact of *BP* removed from the cross section.

10.4.7 Controls for Tax Effects

In testing the assumption of no tax effect in the standard CAPM, Brennen (1972) tested the hypothesis that stock returns should reflect the differential taxation of dividends (ordinary income) and long-term capital gains (taxed at a preferential rate and deferrable). Brennen found that dividend yield was an economically and statistically significant omitted variable in the standard before-tax CAPM and established thereby that the before-tax cross-section of stock returns had a *dividend yield tilt*, a higher before-tax return at a given beta value when a higher fraction of total return is realized as dividends. In the context of the CAPM, the dividend yield tilt has been corroborated in many subsequent studies [see Rosenberg and Rudd (1977, 1982), Rosenberg and Marathe (1979), Blume (1980), Peterson et al. (1985)].[16]

While the *dividend yield tilt* is well-established empirically, it is generally not referred to as a "*return anomaly*" because, like systematic risk factors, the differential taxation of dividends and capital gains should have a well-defined price. In the bond market, it is relatively straight forward to infer an implied marginal tax rate from the yield spread between taxable and tax exempt bonds of the same maturity and default risk (quality rating).

Inferring a tax effect in the stock market is much more complex. In addition to knowing the effective tax rate for dividend income, we also need to know an effective tax rate for return realized as capital gains. However, because gains are taxed when realized, we need to know as well the effective gain realization horizon. The point is that either correcting the cross-section of stock returns for a tax-based dividend-gain trade-off or putting returns on an after-tax basis is complex given uncertainty about both the gain realization horizon and the effective tax rate on capital gains. In addition to tax-based trade-off measurement complexities, there are other non-tax factors that can complicate an empirical assessment of an effective trade-off rate for return realized as dividends including:

- Possibly different prices for a least some systematic risk factors
- A non-tax, non-risk preference on the part of some investors such as endowment funds for return in the form of dividends.

The intent here is not to structure a way to measure and correct the cross-section of stock returns for a tax-based trade-off between return as dividends and return as gains. Rather, the objective here is to note that measuring and correcting for any dividend-gain cross-sectional return effect can be a difficult problem.

When we create a cross-section of portfolios that are matched on both beta and dividend yield, earnings yield, and other gain related variables such as growth, the effect of the combination of controls should be to create a cross-section of control-matched portfolios such that each portfolio should have the same ex ante expectation for dividend return and gain return. By controlling each portfolio for the relative amount of return realized as gains and as dividends, we avoid the need to estimate either the tax rate of the marginal investor or the time horizon for realizing gains. In addition, we avoid at the same time any implicit price arising from any difference in the systematic risk pricing of dividend return vs. gain return. Controlling for the dividend-gain mix is a robust control for possible dividend-gain tax effects and any other return effect arising from differences in the dividend-gain mix in returns.

Of course, the differential taxation of dividends and gains is not the only possible tax-based valuation impact. The tax shield of debt financing is a well-established (but hard to measure) valuation effect. Theories of financial structure say that the mix of debt and equity is pertinent to company value and therefore pertinent

[16] For additional more recent studies, see Fama and French (1988) and Pilotte (2003). For an extensive review of both dividend valuation and dividend policy and extensive references in this area, see Lease et al. (2000).

to explaining the cross-section of stock returns. However, when assumptions of bankruptcy cost, value to financial slack, and revaluation of debt from changes in interest rates are recognized, just what functional form should be used to model the return impact of debt is less than obvious. The question of correct functional specification is further exacerbated by the fact that debt interacts non-linearly with other valuation variables. For instance, debt financing magnifies non-linearly the inherent equity beta as established in Hamada (1972). However, the need to know the functional form of the cross-sectional return dependency on beta is by-passed by requiring that each portfolio in the cross-section has the same average value for the fraction of debt financing.

When the financial structure control is combined with the control of having the same average size, each portfolio should have the same starting mix of debt and equity and therefore the same portfolio-level return response to any financial structure valuation/revaluation effects arising from changes in interest rates. Thus, for sufficiently complete sets of control variables, the time series of control-matched return cross sections should be well-immunized to debt-related revaluations arising from interest rate changes.

10.4.8 Growth

Given conventional wisdom that value and growth are competing investment styles, testing a value hypothesis clearly requires that we isolate value effects from growth. We use two growth controls. The 5-year average sales growth is a measure of past sales growth. In addition, we use an average of *sustainable growth* (grow in total common equity from additions to retained earnings). We use a 3-year average rather than the text book point-in-time measure to smooth variation in additions to retained earnings including fluctuations with the business cycle.

See Sect. 10.8 for a discussion of the problem of predicting growth and the reasons that we have added sustainable growth to past sales growth to ensure better control for differences in future growth.

10.4.9 Profitability and Other Attributes

Profitability ratios include *average operating margin* (operating earnings)/sales, ROA (return on assets), ROI (return on investment), and ROE (return on equity).

Other profitability attributes include *capital intensity* (sales per dollar of total investment) and variations such as sales per dollar of long-term assets.

Capital intensity also serves as a ratio instrument for industry attributes.

10.4.10 Multicollinearity Distortion

When rank-ordering stocks at a point in time on a possible explanatory variable such as the EP or BP ratio and then grouping the rank-ordered stocks into fractile portfolios, the hope is that one is creating observations from a well-defined subsurface.

A strong sufficient condition for having observations belong to a well-defined subsurface is that the ranking variable be independent of each of the other firm-specific explanatory variables. In this case, each explanatory variable should be close to its sample average value in each of the fractile portfolios. The departures from the sample average value should be uncorrelated from portfolio to portfolio. In essence, the departures of each portfolio value from the value associated with the sample average should behave like random errors.

A weak sufficient condition for not having the rank-ordered portfolio observations belong to a well-defined subsurface is the presence of correlation between the dependency variables. For instance, if one ranks on the book-price ratio BP, then well-established cross-sectional correlation with growth, dividends (tax effects), and financial structure mean that the cross-section of fractile portfolios is systematically distorted.[17]

Ranking into fractile portfolios can reduce measurement error and ensure a well-defined conditional subsurface when the other explanatory variables are uncorrelated, a condition rarely satisfied in most practical situations. Multicollinearity is often the limiting factor in using grouping to control measurement error. Stone (2003) shows that rank-ordering stocks in a sample of a thousand-plus stocks into 20 or fewer fractile portfolios can magnify relatively modest sample-level multicollinearity with the rank-ordering variable, e.g., stock sample level correlation coefficients of less than 0.2 can become.8 plus at the decile level.

Fortunately, even in the presence of multicollinearity between a stock return forecast and other return impact variables, one can ensure a well-defined response subsurface by mapping rank-ordered fractile portfolios into portfolios that have the same value for each control. Zero cross-sectional variation in a control variable over the set of forecast-rank-ordered portfolio observations tends to ensure that the realized return observations belong to a well-defined response subsurface (well-defined conditional functional dependency).

[17] If one also uses either a past or a contemporaneous cross-section to infer the market-level return price of BP (e.g., by using high-minus-low portfolio returns), then the distortion of the return response subsurface is compounded even further.

10.5 Mapping Cross Sections of Fractile Portfolios into Control-Matched Forecast-Ranked Cross Sections

10.5.1 The Need to Reassign Securities

Ranking on forecasted return score and grouping into fractile portfolios produces a set of portfolios ordered on the basis of predicted return score. This return cross section will almost certainly have a wide range of forecasted return values. However, each portfolio in the cross section will almost never have the same average values of the control variables listed in Exhibit 10.2. To the extent that the values of the controls fluctuate randomly about their average value over the cross section, their variation is a source of noise (and possibly loss of efficiency) in assessing the cross-sectional dependency of realized returns (and realized return related measures such as standard deviation) on the return forecast score. However, to the extent that a control is correlated with the return forecast score, then systematic variation in the control with the rank-ordering will mean the impact of the control variable is mixed in with any dependency on the return forecast.

Reassigning stocks to produce a new cross section with no portfolio-to-portfolio variation in the control variable means no differential impact on the return cross sections. In addition to possible improvements in efficiency, making all of the portfolios in a cross section have the same portfolio-average value for a correlated control variable means that we now isolate well the dependency of realized returns on the return forecast from any dependency without any differential distortion from the correlated control variable.

To produce a cross-sectional match on any of the control variables, we must reassign stocks. For instance, if we were trying to make each portfolio in the cross section have the same average beta value, we could move a stock with an above-average beta value into a portfolio whose average beta value is below the population average. At the same time, we could shift a stock with a below-average beta value into the above-average portfolio from the below-average portfolio.

Just to produce a match for each portfolio in the cross section on a single explanatory control variable such as beta clearly entails an immense number of possible reassignments of stocks across portfolios. Fortunately, we do not have to use trial-and-error switching of stocks between portfolios to find the best reassignment that produces a cross-sectional match on beta or any other control variable. This reassignment problem can be formulated as a mathematical assignment program (*MAP*). All fractile portfolios should have explanatory controls equal to their population average value.

10.5.2 Notation Summary

P = number of rank-based portfolios in the cross section
p = 1 is the portfolio with the smallest value of the rank-ordering variable

p = P is the portfolio with the largest value of the rank-ordering variable
S = total number of securities being assigned to portfolios
s = security subscript
X_{ps} = the fraction of security s assigned to portfolio p
V = representative control variable
VTARGET = target average value of a representative control variable.[18]

10.5.3 Formulating the MAP: Overview

Given a cross-section of fractile portfolios formed by rank-ordered grouping on the basis of predicted return, the objective of the assignment program is to transform this cross-section of fractile portfolios into an associated control-matched cross section to optimize two complementary attributes of statistical power:

1. Preserving a wide range of well-ordered return forecasts.
2. Preserving within-portfolio homogeneity of forecasted return.

The four constraints are:

1. The portfolio average value of each control variable must equal the population mean
2. The initial size (number of securities) of each portfolio must be preserved
3. Each security must be fully assigned
4. There can be no short sales

The crucial constraints are the control matching restrictions. Preserving initial portfolio size and full use of each security are technical constraints that go with full use of the sample. Prohibiting short sales prevents one return observation from canceling out other return observations. Prohibiting short sales is also consistent with the idea of full use of all sample information in a long-only framework.

10.5.4 The Objective Function

Preserving range and minimizing cross-portfolio mixing are two aspects of statistical power. They are complementary measures in that optimizing one tends to optimize the other. To reflect the relative importance of these two measures, we define a trade-off parameter Φ that defines a relative weighting, where $0 < \Phi < 1$. The trade-off between range and within-portfolio variance can be written as

$$\text{OBJECTIVE} = \Phi\,[\text{RANGE}] - (1 - \Phi)\,[\text{WITHIN-PORTFOLIO VARIANCES}]\,.$$

[18] In this study, the target average value is always the ex ante sample average value.

For each portfolio in the cross section, the within portfolio variance is the portfolio weighted squared deviation of return forecast score from the portfolio mean forecast return score. It is a quadratic function. Thus, minimizing within-portfolio variance, actually minimizing a sum of within-portfolio variances over the cross section, means a quadratic objective function.

In this study in which we assess month-to-month return cross sections in each of the 456 months of 1967–2004, we impose progressively more complete sets of control variables in each month. Obtaining 15 or more control-matched cross sections in 456 month means solving more than 6,700 optimization runs. Solving this many quadratic programs would be a computational challenge. However, just as one can approximate well the mean–variance portfolio optimization of Markowitz (1952, 1959) by solving an associated linear programming (LP) approximation to the quadratic program [see for instance, Sharpe (1963, 1967, 1971), Stone (1973) and Elton et al. (1979)], we can approximate the control-matching quadratic optimization by an associated LP objective function.

Proceeding directly with the MAP formulation using an LP approximation objective function, we write the LP approximation objective function as

$$\text{Maximize : LP OBJECTIVE} = \Phi\,[\text{RANGE}] - (1 - \Phi)\,[\text{SHIFTING}]\,. \qquad (10.4)$$

The measure SHIFTING is the approximation to variance minimization that we now define.[19] Let D_{ps} be the squared difference in the numerical rank between portfolio p and the natural portfolio rank of security s in the initial rank-order partitioning into fractile portfolios. The set of D_{ps} can be summarized by a symmetric $\mathrm{P} \times \mathrm{S}$ matrix. Squaring the difference in numerical rank means that large shifts are much worse than small ones.

If $\mathrm{FS_s}$ denotes the value of the forecast score for stock s, then the objective function above can be written in terms of assignment variables as:

$$\text{Maximize } \Phi[\Sigma_s X_{Ps} FS_s - \Sigma_s X_{1s} FS_s\,] - (1 - \Phi)]\, \Sigma_p \Sigma_s X_{ps} D_{ps}] \qquad (10.5)$$

The MAP can be solved for a range of trade-off values by varying Φ from zero to 1. In the results reported in this paper, the value of the trade-off parameter Φ is

[19] It is intuitive that minimizing the amount and distance of cross portfolio shifting tends to preserve the original within-portfolio forecast distribution including within-portfolio variances. The substance of this approximation is to use portfolio rank-order distance as a substitute for actual return forecast differences. Since we map each return forecast into a near uniform distribution on the (0, 1) interval, we tend to ensure the validity of this approximation. Moreover, in the regression assessments summarized in Exhibits 7, 8, 9, and 10 of how well the longrun realized cross sections of returns, Sharpe ratios, standard deviations, and skewness coefficients are explained by the return forecasts, we show that using portfolio rank and using the return forecast score as explanatory variables produce almost identical regression fits, which means that the portfolio rank-order number and the normalized to (0, 1) return forecast scores have essentially the same (to within a linear scaling) regression errors. This consistency tends to validate for at least the longrun average results the validity of using rank-order differences as approximations for differences in the normalized forecast scores.

0.25. However, experience shows that the solutions are robust to variation in the trade-off Φ. The reason for the robustness is that these two attributes of statistical power are complementary objectives. Minimizing cross-fractile shifting generally preserves most of the range as well as the distribution of return forecast scores in the starting fractile portfolios.

10.5.5 Equal Value Constraint for Each Control Variable

Let V_{sj} denote the security s value of a representative control variable j. Let $VTARGET_j$ denote the target value of this representative control variable for all P portfolios in the cross section. The representative control constraint can be expressed as

$$\sum_s X_{ps} V_{sj} = V\text{TARGET}_j, \quad p = 1, \ldots, P-1. \tag{10.6}$$

10.5.6 Security Usage and Short Sales: Technical Constraints

We impose two generic data usage constraints. The first says that each security must be fully assigned to one or more portfolios, i.e.:

$$\sum_p X_{ps} = 1, \quad s = 1, \ldots, S \tag{10.7}$$

The second security assignment constraint keeps the number of securities in each matched portfolio the same as the number of securities in the corresponding fractile of the starting rank-order partitioning. Let F_p denote the number of securities in fractile p. Then this restriction is

$$\sum_s X_{ps} = F_p, \quad p = 1, \ldots, P \tag{10.8}$$

The no short-sale restriction and the natural limitation that no security can be used more than once requires:

$$0 \leq X_{ps} \leq 1 \tag{10.9}$$

10.5.7 Synthesis of Reassignment Optimizer

The substance of the reassignment process is well understood by knowing input and output. The *input* is a cross-section formed by ranking stocks into fractile portfolios, which is the focus of most cross-sectional return analyses in past work on cross-sectional return dependencies.

The *output* is a cross-section of fractile portfolios that are matched on a specified set of controls variables. The MAP finds an optimal reassignment of stocks that transforms the input rank-ordered cross section into a new cross section that is matched on the portfolio average values of each control variable.

Optimization arises in finding the particular reassignment that optimizes a trade-off between preserving the widest possible range of well-ordered portfolio values of forecasted return and also ensuring preservation of within-portfolio homogeneity of forecasted return. For the LP transformation used to obtain control-matched cross sections, the linear objective function is a trade-off between preserving as much as possible a measure of cross-portfolio range while minimizing the shifting of stocks away from the original rank-order portfolio. Even though we have used an LP to construct a control-matched cross section, any approximation is in the objective function and not in meeting any constraints. The constraints requiring that every portfolio in the cross section has an exact match on each control variable are met exactly in each month for every control variable in every portfolio.

Given the sample of stocks with variable values for each stock in that time period, once we pick a number of portfolios P in the cross section and select a set of control variables, the transformation of the rank-ordered cross section into the control-matched cross section is defined by the optimization program. The mapping from the rank-ordered cross section into the control-matched cross section is objective in the sense that the researcher (or researchers) exercises no discretion in how stocks are reassigned. The data and the MAP determine the output cross section.

10.6 Data Sample and Empirical Procedure

10.6.1 Sample Formation; Growth in Sample Size for 1968–2004 Test Period

The stocks used to form portfolios are all non-financial common stocks in the CRSP–COMPUSTAT intersections that have been listed at least 5 years at the time of portfolio formation for which all return and all financial statement data are available.

Because of the sparseness of the Compustat database in the 1963–1967 start-up period required for variables such as 5-year sales growth, there are only 324 companies in the first forecast month, January 1968. Exhibit 10.3 provides a year-by-year summary of the number of stocks in the forecast sample in January of each year. Exhibit 10.3 shows that the forecast sample size grows rapidly. From 1971 on, there are more than 900 companies in the forecast sample growing to more than 2,000 companies by 1995. The fact that the sample size shows little growth from the 2,003

Exhibit 10.3 Sample Size: Number of Stocks in Sample in January of Each Year

Year	#Stocks	Year	#Stocks
1967	198	1986	1660
1968	324	1987	1632
1969	422	1988	1580
1970	564	1989	1621
1971	901	1990	1644
1972	966	1991	1671
1973	1058	1992	1742
1974	1108	1993	1845
1975	1037	1994	1921
1976	1329	1995	2003
1977	1495	1996	2057
1978	1651	1997	2193
1979	1701	1998	2238
1980	1703	1999	2331
1981	1757	2000	2284
1982	1734	2001	2256
1983	1698	2002	2305
1984	1714	2003	2318
1985	1676	2004	2238

companies in January 1995 to the 2,238 in January 2004 indicates that the large number of new IPOs from the mid-1990s on is not producing an increase in the number of sample companies. The fact that our sample does not exhibit the same growth as the cross time increase in publicly listed companies shows that the combination of data availability and minimum book value restrictions mean that we are studying primarily larger more mature companies.

10.6.2 Number of Fractile Portfolios

Conventional practice in cross-sectional return dependency assessments has been to form deciles and more recently only quintiles. There are several reasons for using 30 fractile portfolios rather than forming deciles or even quintiles as done in some recent studies. Using a larger number pertains to the power-efficiency trade-off. First, grouping (averaging) tends to lose information, especially in the tails of the distribution while most of the efficiency benefits of measurement error and omitted variable diversification are accomplished with 20 or fewer stocks in a fractile. Second, to do regression fits to the return and standard deviation cross sections, more portfolio observations are clearly preferred to less with at least 20 being desirable. Third, Stone (2003) shows that very low levels of sample correlation coefficients are magnified nonlinearly when one groups by rank-ordering on any one variable. The loss of power from colinearity confounding accelerates rapidly when there are 20 or fewer portfolios in the cross-section of fractile portfolios. Fourth, given that one can group together adjacent portfolios in a control-matched return cross section (as done here in looking at both the top quintile (top six portfolios grouped) and bottom quintile (bottom six portfolios grouped), one should probably error on the side of too many portfolios in a cross section rather than too few.

10.6.3 Stepwise Imposition of Control Constraints

Given a return forecast for each stock in the sample and a rank-ordering into portfolio fractiles, we have a return cross-section with no controls. We input this cross section to the MAP with all reasonable controls and obtain thereby a well-isolated cross-sectional dependency of realized return on return forecast score to the extent that there are no omitted controls. However, rather than going from no controls to a full set of controls in one step, it is useful to add controls in a stepwise fashion. Adding controls in a stepwise fashion as we do in our empirical tests enables us to explore how the initial rank-ordered cross-section changes as we remove the effect of a control variable or combination of control variables. This stepwise exploration of how the return dependency changes with changes in combinations of control variables is generally very informative.

Given that a primary concern is risk-corrected, tax-corrected return performance, it is logical to begin assessing the impact of individual risk controls such as beta, size, and the book-price ratio. Then, one can evaluate the combination of all three risk controls together. Many combinations of controls are possible. Exhibit 10.4 summarizes the order in which we report stepwise imposition of control constraints. For instance, we add our three tax controls as a group and then as a group with the three Fama–French risk controls. The empirical work reported in Sects. 10.8–10.12 shows that the three tax controls have a significant impact *after* imposing the

Exhibit 10.4 A Summary of the Sets of Control Variables in the Order of their Stepwise Imposition in Exhibits 10.5 to 10.12

Order	Control variables in each constraint set
A	**No Control Constraints (rank-ordering only)**
B	**β (Beta)**
C	**S (Size)**
D	**BP (Book-to-Market)**
E	**β, S, BP**
F	**FL, EP, DP**
G	**β, S, BP, FL, EP, DP**
H	**β, S, DP, FL, Sag5, Sug3**
I	**β, S, BP, FL, EP, DP, Sag5, Sug3, ROE**
J	**β, S, BP, FL, EP, DP, Sag5, Sug3, ROI**
K	**β, S, BP, FL, EP, DP, Sag5, Sug3, SI**
L	**β, S, BP, FL, EP, DP, Sag5, Sug3, ROE, SI**
M	**β, S, BP, FL, EP, DP, SAG5, SUG3, ROE, SI, CP, SP**

standard risk controls. Thus, from stepwise imposition of controls, we learn that the three standard risk controls alone do not ensure a well-isolated return-forecast response subsurface. We also learn that the three tax controls have both an economically and statistically significant impact on the return cross-section.

10.6.4 Assessing the Impact of Individual Model Variables

Since the book-price ratio is a model variable, its use as a control means that any contribution of BP to return performance is removed from the cross-section. The same comment can be made for the EP ratio once we impose tax controls. The key point here is that we can isolate return response from not only non-model control variables but also forecast model variables. Thus, it is possible to assess the relative contribution to both return and risk of individual forecast variables or forecast variable combinations by making them into controls.

10.7 Output Summary: The Cross Sections of Average Realized Returns, Realized Standard Deviations, Realized Skewness Coefficients, and Realized Sharpe Ratios for Selected Sets of Control Variables

10.7.1 Procedure Summary

The following summarizes the steps followed in each study month for each set of control constraints to form the cross-section of 30 control-matched portfolios.

Step 1: Generate Return Forecast. Using the return forecast procedure outlined in Sect. 10.3 update the return forecast model and generate a return forecast score for every stock in the sample at that time.
Step 2: Rank-Ordering. Rank-order all sample stocks on their return forecast score.
Step 3: Form Fractile Portfolios. Form 30 fractile portfolios by grouping adjacently ranked stocks.
Step 4: Transform Fractile Portfolios into Control-Matched Portfolios. Use the MAP (mathematical assignment program) to reassign stocks in the starting set of 30 forecast ranked portfolios to produce a new portfolio cross-section matched on specified control variables.
Step 5: Stepwise Imposition of Controls. Repeat step 4 for each of the control sets listed in Exhibit 10.4.
Step 6: Input to Measurement and Statistical Testing. Store the data on each return cross section for each set of control restrictions for input to the cross time performance assessments.

10.7.2 Aggregation of the Time Series of Cross Sections

For each set of variable controls, the output of the return cross-section formation summarized above is a time series of 30 control-matched portfolio cross sections that have a wide range of well-ordered return forecast scores. In a multivariate extension of the Fama–MacBeth (1973) CAPM-focused time series of return cross sections,[20] we can assess the cross time realized returns and other performance data for each portfolio rank in the cross section for each set of variable controls for pertinent subperiods and the overall study time.

The following summarizes the cross time measurements for each portfolio in the rank-ordered cross section.

1. Average realized returns
2. Realized cross-time portfolio standard deviation
3. Realized cross-time portfolio skewness coefficients
4. Realized Sharpe ratios.

10.7.3 An Empirical Overview of the Empirical Tests

The starting point for input to the MAP is a cross-section of stocks rank-ordered on forecasted return score and partitioned into 30 fractile portfolios. While one could simply go from this input to a cross-section matched on all the controls, a researcher can learn more about the impact of controls on the cross section by imposing the control restrictions in a stepwise fashion. The stepwise imposition of controls outlined here is a progressive exploration of how the cross-section of realized returns depends jointly on the return forecast score used to form the cross-section of fractile portfolios and other return impact variables that may be correlated or at least partially correlated with the return forecast score. The process of stepwise imposition of progressively more complete control sets is a process of moving toward a progressively more well-defined (well-isolated) return-forecast response subsurface.

Given limited space, it is impossible to report all possible combinations of control constraints. As noted previously, Exhibit 10.4 summarizes sets of control constraints in the order that we report empirical results. The following is an overview of the exhibits summarizing empirical results.

1. *Exhibits 10.5A through 10.5L* provide a portfolio-by-portfolio tabulation and graphing of long-run average returns, realized cross-time standard deviations, realized cross-time skewness coefficients (not graphed), and realized long-run

[20] See also Chan et al. (1991) and Lakonishok et al. (1994) for multivariate pooling of time series cross sections.

Exhibit 10.5A Annualized Monthly Average Returns: No Control Constraints

Portfolio	Forecast Score	Annualized Risky Return %	Risky Return Stdev %	Skewness	Sharpe Ratio
1	2.98	10.29	8.96	0.802	0.096
2	3.90	6.22	9.11	0.279	0.057
3	6.99	8.84	7.96	0.065	0.093
4	11.70	7.72	7.01	−0.231	0.092
5	14.38	6.10	6.81	−0.300	0.075
6	18.02	8.89	6.41	−0.543	0.116
7	21.51	7.94	6.13	−0.509	0.108
8	24.98	7.75	6.09	−0.519	0.106
9	28.05	9.18	5.76	−0.682	0.133
10	31.50	7.85	5.68	−0.593	0.115
11	34.84	6.48	5.54	−0.714	0.097
12	37.98	7.96	5.56	−0.606	0.119
13	41.04	9.92	5.48	−0.591	0.151
14	44.48	9.20	5.43	−0.528	0.141
15	47.98	9.39	5.39	−0.503	0.145
16	51.03	9.42	5.35	−0.384	0.147
17	54.45	8.79	5.35	−0.417	0.137
18	57.89	9.93	5.54	−0.213	0.149
19	60.96	9.48	5.50	−0.209	0.144
20	64.44	11.74	5.56	−0.241	0.176
21	67.87	9.97	5.69	−0.077	0.146
22	70.89	10.04	5.78	0.034	0.145
23	73.96	10.36	6.08	0.462	0.142
24	77.43	12.39	6.35	0.287	0.163
25	80.87	11.95	6.49	0.530	0.154
26	84.41	12.93	6.95	0.421	0.155
27	87.06	14.29	7.35	1.061	0.162
28	91.08	13.03	8.26	0.993	0.131
29	94.29	15.50	9.57	1.271	0.135
30	96.59	19.12	10.62	2.308	0.150

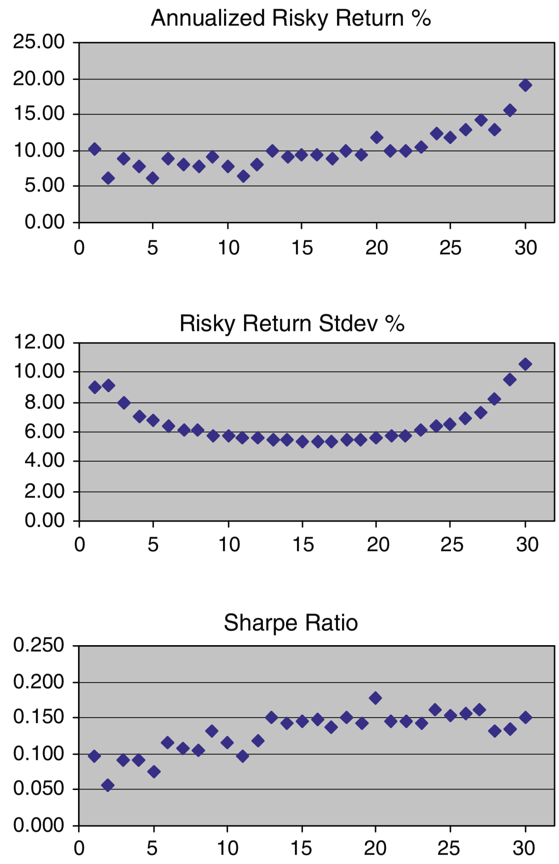


Exhibit 10.5B Annualized Monthly Average Returns Controls: Beta Only

Portfolio	Forecast Score	Annualized Risky Return %	Risky Return Stdev %	Skewness	Sharpe Ratio
1	4.21	7.77	7.36	−0.248	0.088
2	5.20	5.74	7.59	0.080	0.063
3	8.54	9.37	7.14	−0.447	0.109
4	13.52	7.72	6.22	−0.617	0.103
5	16.19	7.64	6.40	−0.636	0.099
6	19.69	9.61	5.83	−0.565	0.137
7	23.12	7.73	5.98	−0.612	0.108
8	26.27	7.60	5.81	−0.477	0.109
9	29.14	7.56	5.78	−0.507	0.109
10	32.34	9.83	5.68	−0.629	0.144
11	35.49	7.94	5.56	−0.400	0.119
12	38.51	7.73	5.77	−0.541	0.112
13	41.44	7.68	5.61	−0.567	0.114
14	44.72	8.70	5.64	−0.402	0.128
15	48.02	8.08	5.54	−0.484	0.121
16	50.97	9.39	5.65	−0.404	0.139
17	54.24	8.85	5.59	−0.290	0.132
18	57.51	9.04	5.76	−0.034	0.131
19	60.47	9.17	5.75	0.197	0.133
20	63.79	11.05	5.84	−0.012	0.158
21	67.07	10.47	5.78	−0.180	0.151
22	69.96	10.37	5.92	0.149	0.146
23	72.92	10.89	5.85	−0.010	0.155
24	76.15	12.23	6.09	0.042	0.167
25	79.65	12.76	6.07	0.287	0.175
26	83.18	11.10	6.50	0.632	0.142
27	85.85	13.27	6.78	0.247	0.163
28	89.97	15.10	7.65	1.263	0.164
29	93.48	15.11	8.68	1.483	0.145
30	95.94	19.89	9.55	1.585	0.173

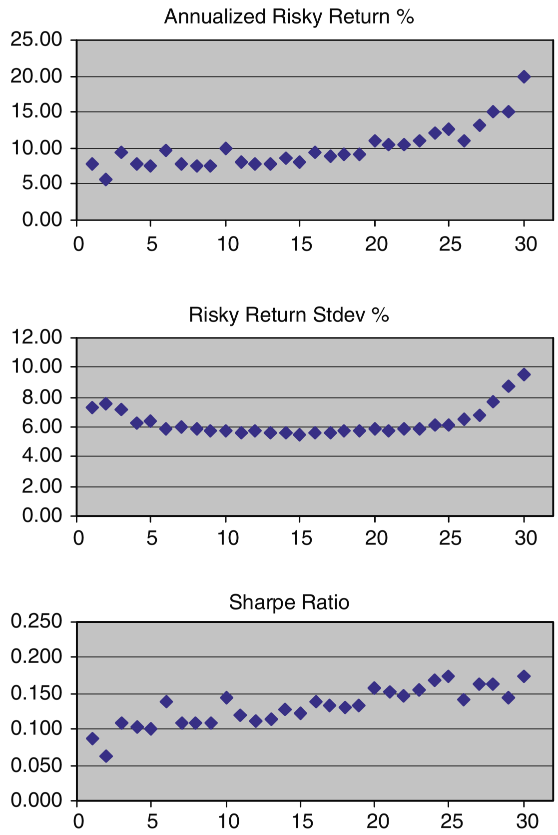


Exhibit 10.5C Annualized Monthly Average Returns Controls: Size Only

Portfolio	Forecast Score	Annualized Risky Return %	Risky Return Stdev %	Skewness	Sharpe Ratio
1	3.28	10.31	8.90	0.914	0.097
2	4.22	6.77	9.08	0.159	0.062
3	7.31	7.73	7.93	0.137	0.081
4	12.00	7.91	7.03	−0.313	0.094
5	14.65	6.35	6.85	−0.200	0.077
6	18.28	8.70	6.33	−0.556	0.115
7	21.86	8.06	6.06	−0.621	0.111
8	25.20	7.79	6.10	−0.538	0.106
9	28.29	9.07	5.84	−0.602	0.130
10	31.67	7.09	5.61	−0.704	0.105
11	34.97	6.80	5.54	−0.640	0.102
12	38.11	8.49	5.54	−0.537	0.128
13	41.16	9.49	5.47	−0.581	0.145
14	44.51	9.05	5.35	−0.604	0.141
15	47.99	9.34	5.44	−0.413	0.143
16	51.09	9.30	5.33	−0.375	0.145
17	54.40	8.84	5.36	−0.478	0.137
18	57.82	10.07	5.58	−0.279	0.150
19	60.90	9.79	5.42	−0.198	0.150
20	64.26	11.35	5.51	−0.196	0.172
21	67.67	10.17	5.62	−0.035	0.151
22	70.68	10.91	5.95	0.266	0.153
23	73.75	10.31	5.96	0.189	0.144
24	77.07	12.86	6.31	0.340	0.170
25	80.63	11.73	6.52	0.322	0.150
26	84.15	12.78	6.96	0.514	0.153
27	86.80	14.43	7.29	1.045	0.165
28	90.80	13.24	8.21	1.017	0.134
29	94.01	15.06	9.58	1.261	0.131
30	96.28	18.87	10.55	2.329	0.149

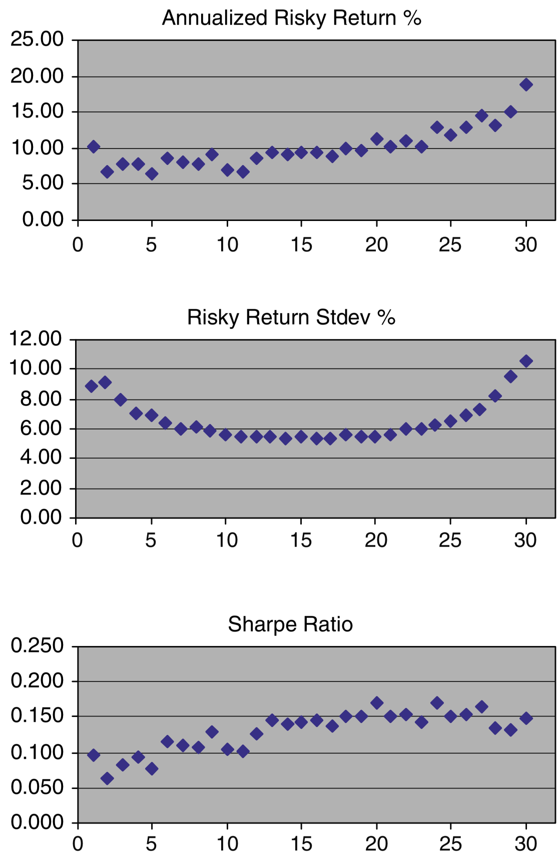
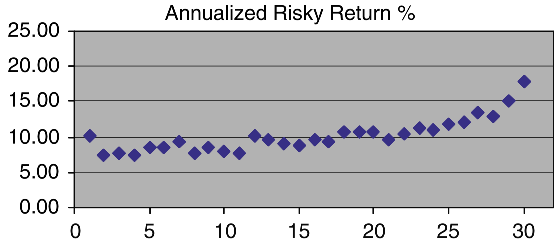
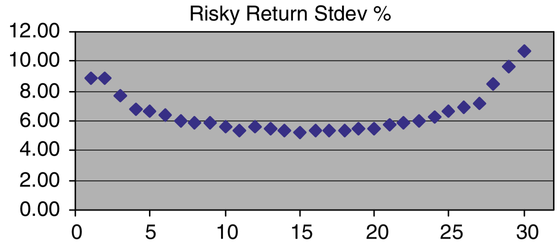


Exhibit 10.5D Annualized Monthly Average Returns Controls: BP (Book to Market) Only

Portfolio	Forecast Score	Annualized Risky Return %	Risky Return Stdev %	Skewness	Sharpe Ratio
1	4.03	10.27	8.83	0.805	0.097
2	5.47	7.47	8.89	0.276	0.070
3	9.10	7.66	7.64	−0.064	0.083
4	14.47	7.46	6.83	−0.460	0.091
5	17.26	8.51	6.71	−0.495	0.106
6	21.01	8.60	6.45	−0.292	0.111
7	24.59	9.23	5.94	−0.657	0.129
8	27.75	7.76	5.88	−0.745	0.110
9	30.61	8.45	5.87	−0.772	0.120
10	33.74	8.09	5.67	−0.515	0.119
11	36.75	7.70	5.38	−0.824	0.119
12	39.56	10.29	5.55	−0.547	0.154
13	42.27	9.68	5.51	−0.690	0.146
14	45.27	8.96	5.38	−0.569	0.139
15	48.26	8.93	5.25	−0.594	0.142
16	50.89	9.74	5.35	−0.425	0.152
17	53.84	9.24	5.33	−0.381	0.144
18	56.82	10.67	5.37	−0.310	0.166
19	59.50	10.67	5.51	−0.077	0.161
20	62.58	10.70	5.49	−0.173	0.162
21	65.68	9.64	5.72	−0.105	0.140
22	68.44	10.50	5.83	−0.038	0.150
23	71.36	11.25	6.02	0.177	0.156
24	74.60	10.88	6.30	0.220	0.144
25	78.16	11.75	6.70	0.565	0.146
26	81.86	12.03	6.92	0.373	0.145
27	84.69	13.33	7.15	0.707	0.155
28	89.21	12.84	8.44	1.469	0.127
29	92.95	15.15	9.62	1.342	0.131
30	95.83	17.85	10.65	2.186	0.140





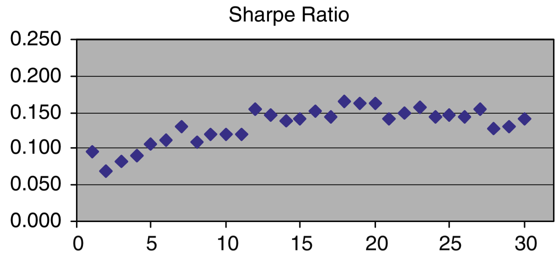


Exhibit 10.5E Annualized Monthly Average Returns Controls: Beta, Size, BP

Portfolio	Forecast Score	Annualized Risky Return %	Risky Return Stdev %	Skewness	Sharpe Ratio
1	5.40	9.88	7.32	−0.405	0.112
2	6.84	4.51	7.17	−0.141	0.052
3	10.62	10.20	7.07	−0.214	0.120
4	16.10	9.17	6.18	−0.591	0.124
5	18.89	8.26	6.23	−0.653	0.110
6	22.45	9.78	5.85	−0.785	0.139
7	25.86	7.53	5.77	−0.668	0.109
8	28.82	9.17	5.91	−0.468	0.129
9	31.55	8.19	5.70	−0.604	0.120
10	34.54	8.20	5.59	−0.661	0.122
11	37.39	8.68	5.74	−0.544	0.126
12	40.09	8.14	5.54	−0.629	0.122
13	42.69	8.80	5.67	−0.589	0.129
14	45.58	8.75	5.60	−0.576	0.130
15	48.49	8.13	5.57	−0.460	0.122
16	51.06	8.07	5.55	−0.306	0.121
17	53.89	8.73	5.51	−0.263	0.132
18	56.76	9.98	5.68	−0.142	0.146
19	59.33	9.73	5.64	−0.099	0.144
20	62.21	10.21	5.76	−0.030	0.148
21	65.12	10.73	5.89	0.007	0.152
22	67.74	11.03	5.88	0.159	0.156
23	70.46	11.47	6.04	0.140	0.158
24	73.51	10.69	5.95	−0.102	0.150
25	76.92	12.70	6.26	0.259	0.169
26	80.49	11.67	6.46	0.377	0.151
27	83.25	12.76	6.80	0.575	0.156
28	87.77	14.59	7.48	0.985	0.163
29	91.80	15.14	8.67	1.230	0.146
30	94.82	18.56	9.34	1.617	0.166

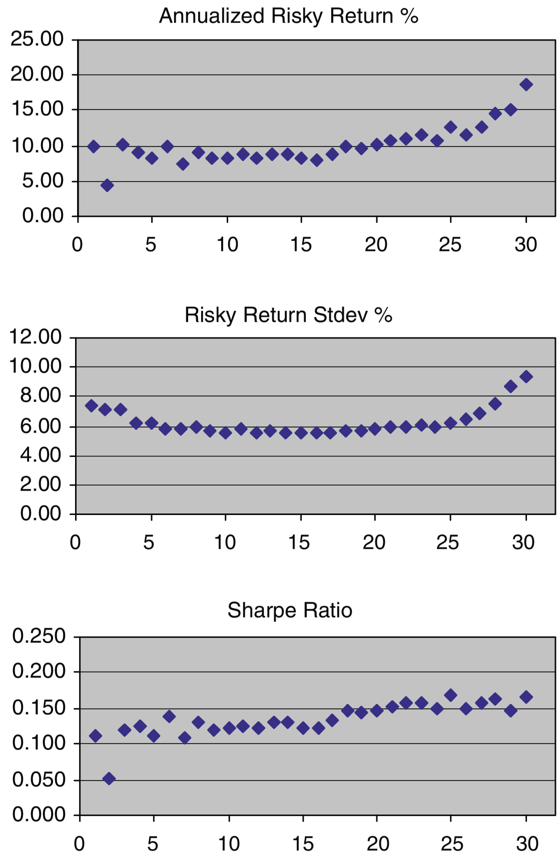


Exhibit 10.5F Annualized Monthly Average Returns Controls: FL, EP, DP

Portfolio	Forecast Score	Annualized Risky Return %	Risky Return Stdev %	Skewness	Sharpe Ratio
1	8.95	1.73	7.82	1.354	0.018
2	9.69	−1.02	7.46	0.041	−0.011
3	13.41	1.72	6.86	−0.206	0.021
4	18.16	2.88	6.39	−0.453	0.037
5	20.43	4.34	6.28	−0.143	0.058
6	23.24	5.13	5.93	−0.327	0.072
7	26.06	5.55	5.56	−0.743	0.083
8	28.46	7.15	5.69	−0.426	0.105
9	30.74	8.24	5.27	−0.715	0.130
10	33.35	7.95	5.30	−0.742	0.125
11	35.95	9.24	5.45	−0.580	0.141
12	38.51	8.80	5.42	−0.644	0.135
13	41.05	8.31	5.52	−0.584	0.125
14	43.99	10.07	5.49	−0.492	0.153
15	47.04	8.80	5.56	−0.367	0.132
16	49.81	9.45	5.62	−0.298	0.140
17	52.92	11.03	5.47	−0.357	0.168
18	56.10	10.69	5.70	−0.303	0.156
19	59.02	11.08	5.82	−0.197	0.159
20	62.36	12.36	5.84	−0.301	0.176
21	65.67	12.16	6.03	0.047	0.168
22	68.59	11.79	6.17	0.015	0.159
23	71.64	12.56	6.20	0.336	0.169
24	74.96	12.73	6.31	0.120	0.168
25	78.56	13.12	6.72	0.421	0.163
26	82.21	13.08	6.90	0.304	0.158
27	84.90	15.60	7.34	0.729	0.177
28	89.24	16.74	8.03	0.865	0.174
29	92.89	17.19	9.16	1.470	0.156
30	95.46	21.33	10.23	2.166	0.174

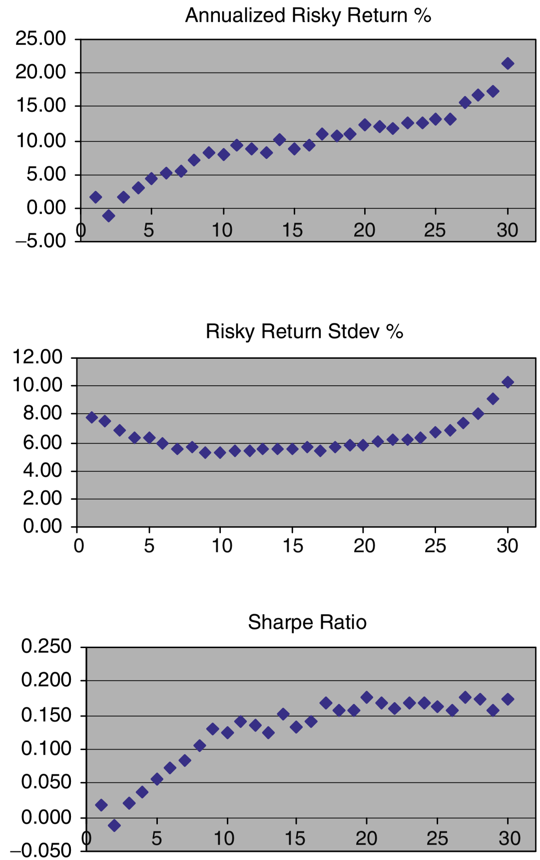
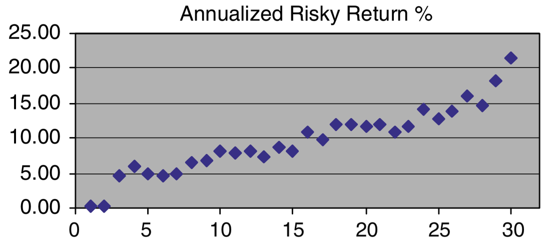
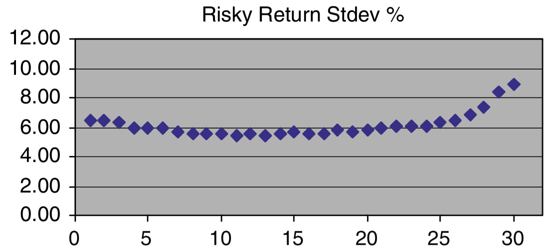


Exhibit 10.5G Annualized Monthly Average Returns Controls: Beta, Size, BP, FL, EP, DP

Portfolio	Forecast Score	Annualized Risky Return %	Risky Return Stdev %	Skewness	Sharpe Ratio
1	10.42	0.30	6.50	−0.420	0.004
2	11.42	0.15	6.45	−0.393	0.002
3	15.46	4.50	6.35	−0.600	0.059
4	20.52	6.09	5.89	−0.737	0.086
5	22.84	4.93	5.98	−0.495	0.069
6	25.76	4.61	5.87	−0.728	0.065
7	28.62	4.90	5.67	−0.642	0.072
8	31.02	6.47	5.57	−0.729	0.097
9	33.27	6.75	5.59	−0.490	0.100
10	35.77	8.13	5.60	−0.638	0.121
11	38.14	7.96	5.47	−0.645	0.121
12	40.46	8.12	5.49	−0.587	0.123
13	42.74	7.33	5.47	−0.607	0.112
14	45.30	8.78	5.59	−0.522	0.131
15	47.90	8.26	5.70	−0.377	0.121
16	50.26	10.91	5.51	−0.407	0.165
17	52.94	9.86	5.60	−0.277	0.147
18	55.63	11.84	5.77	−0.077	0.171
19	58.13	12.02	5.72	−0.269	0.175
20	60.96	11.56	5.83	−0.010	0.165
21	63.80	11.84	5.92	0.047	0.167
22	66.32	10.84	6.02	0.019	0.150
23	68.98	11.79	6.07	0.147	0.162
24	71.98	14.23	6.05	0.069	0.196
25	75.37	12.79	6.38	0.144	0.167
26	78.92	13.99	6.39	0.121	0.182
27	81.66	16.00	6.90	0.604	0.193
28	86.26	14.75	7.38	0.515	0.166
29	90.47	18.25	8.38	0.987	0.181
30	93.65	21.40	8.92	1.789	0.200





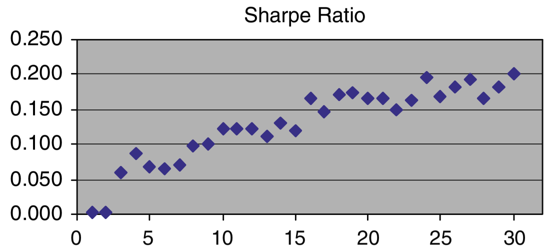


Exhibit 10.5H Annualized Monthly Average Returns Controls: Beta, Size, DP, FL, Sag5, Sug3

Portfolio	Forecast Score	Annualized Risky Return %	Risky Return Stdev %	Skewness	Sharpe Ratio
1	8.59	2.03	6.82	−0.392	0.025
2	9.44	−1.68	6.58	−0.375	−0.021
3	13.12	4.34	6.49	−0.391	0.056
4	18.09	2.84	6.04	−0.716	0.039
5	20.35	3.80	6.00	−0.560	0.053
6	23.29	4.82	5.80	−0.528	0.069
7	26.22	5.59	5.72	−0.594	0.081
8	28.71	5.37	5.74	−0.548	0.078
9	31.08	6.66	5.59	−0.665	0.099
10	33.75	7.61	5.53	−0.592	0.115
11	36.41	6.32	5.45	−0.580	0.097
12	38.98	7.66	5.61	−0.394	0.114
13	41.52	7.82	5.59	−0.559	0.116
14	44.40	8.11	5.45	−0.549	0.124
15	47.37	9.57	5.67	−0.493	0.141
16	50.08	8.83	5.70	−0.307	0.129
17	53.14	10.89	5.57	−0.335	0.163
18	56.26	10.81	5.76	−0.252	0.156
19	59.07	11.43	5.86	−0.028	0.162
20	62.28	12.21	5.80	−0.095	0.175
21	65.44	12.58	6.00	−0.053	0.175
22	68.26	12.65	6.07	−0.165	0.174
23	71.19	11.94	6.07	−0.029	0.164
24	74.40	14.62	6.09	0.219	0.200
25	77.88	12.62	6.29	0.177	0.167
26	81.42	15.74	6.69	0.494	0.196
27	84.05	15.25	6.78	0.419	0.187
28	88.32	17.17	7.68	1.131	0.186
29	91.97	17.70	8.13	1.469	0.181
30	94.59	25.92	9.43	1.552	0.229

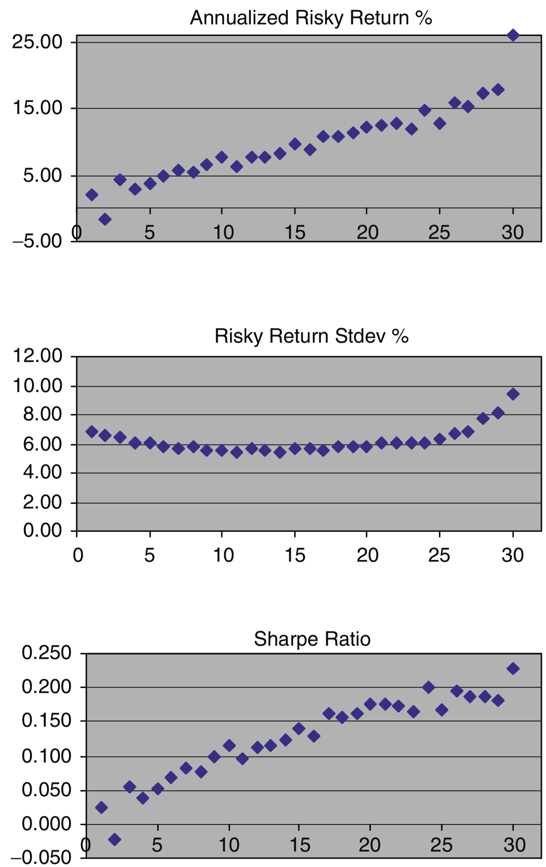
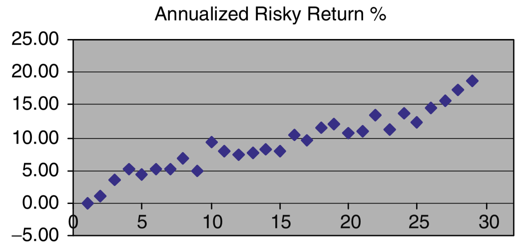
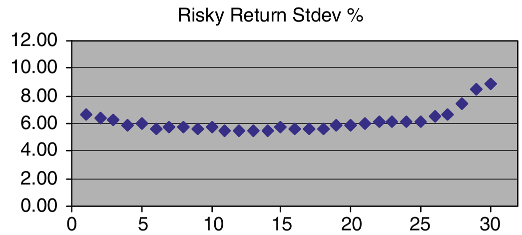


Exhibit 10.5I Annualized Monthly Average Returns Controls: Beta, Size, BP, FL, EP, DP, Sag5, Sug3, ROE

Portfolio	Forecast Score	Annualized Risky Return %	Risky Return Stdev %	Skewness	Sharpe Ratio
1	10.84	−0.18	6.61	−0.511	−0.002
2	11.79	0.93	6.43	−0.459	0.012
3	15.81	3.65	6.24	−0.501	0.049
4	20.79	5.11	5.84	−0.685	0.073
5	23.10	4.42	6.01	−0.628	0.061
6	26.02	5.06	5.65	−0.700	0.075
7	28.88	5.26	5.74	−0.650	0.076
8	31.29	6.79	5.69	−0.600	0.100
9	33.53	4.96	5.57	−0.553	0.074
10	36.03	9.31	5.70	−0.595	0.136
11	38.41	7.96	5.43	−0.672	0.122
12	40.72	7.42	5.50	−0.521	0.112
13	43.00	7.55	5.47	−0.602	0.115
14	45.54	8.18	5.54	−0.637	0.123
15	48.10	7.84	5.71	−0.403	0.114
16	50.44	10.32	5.64	−0.267	0.152
17	53.09	9.54	5.57	−0.356	0.143
18	55.79	11.54	5.64	−0.273	0.171
19	58.24	12.08	5.83	−0.176	0.173
20	61.00	10.75	5.81	−0.051	0.154
21	63.77	11.00	6.00	0.068	0.153
22	66.23	13.48	6.07	0.210	0.185
23	68.80	11.12	6.09	0.059	0.152
24	71.69	13.62	6.08	0.156	0.187
25	74.98	12.32	6.16	−0.058	0.167
26	78.39	14.41	6.55	0.406	0.183
27	80.98	15.59	6.70	0.401	0.194
28	85.42	17.23	7.44	0.546	0.193
29	89.52	18.57	8.49	1.595	0.182
30	92.75	26.49	8.92	1.336	0.248





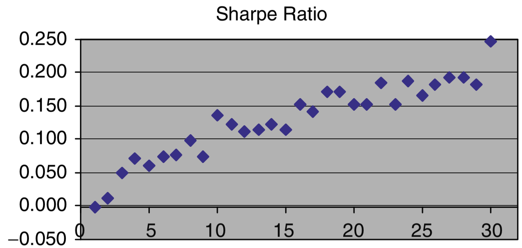


Exhibit 10.5J Annualized Monthly Average Returns Controls: Beta, Size, BP, FL, EP, DP, Sag5, Sug3, ROI

Portfolio	Forecast Score	Annualized Risky Return %	Risky Return Stdev %	Skewness	Sharpe Ratio
1	10.86	−0.28	6.56	−0.546	−0.004
2	11.83	0.23	6.43	−0.526	0.003
3	15.84	4.56	6.24	−0.497	0.061
4	20.82	4.87	5.85	−0.721	0.069
5	23.14	3.65	5.96	−0.555	0.051
6	26.07	5.72	5.83	−0.735	0.082
7	28.94	6.24	5.62	−0.655	0.093
8	31.37	6.38	5.84	−0.512	0.091
9	33.62	6.15	5.59	−0.654	0.092
10	36.11	8.02	5.56	−0.529	0.120
11	38.49	7.78	5.49	−0.654	0.118
12	40.79	7.82	5.58	−0.550	0.117
13	43.05	7.84	5.43	−0.547	0.120
14	45.58	7.40	5.54	−0.651	0.111
15	48.16	9.38	5.68	−0.452	0.138
16	50.48	10.11	5.69	−0.322	0.148
17	53.10	9.44	5.58	−0.319	0.141
18	55.77	11.88	5.68	−0.136	0.174
19	58.20	12.38	5.79	−0.272	0.178
20	60.96	11.56	5.81	0.005	0.166
21	63.68	10.48	6.02	0.061	0.145
22	66.13	12.58	6.06	0.147	0.173
23	68.71	10.87	6.11	0.086	0.148
24	71.60	13.48	6.09	0.034	0.184
25	74.89	13.46	6.28	0.013	0.178
26	78.29	13.58	6.50	0.433	0.174
27	80.90	16.19	6.80	0.477	0.199
28	85.34	16.94	7.36	0.719	0.192
29	89.47	19.85	8.26	1.089	0.200
30	92.74	24.67	8.91	1.580	0.231

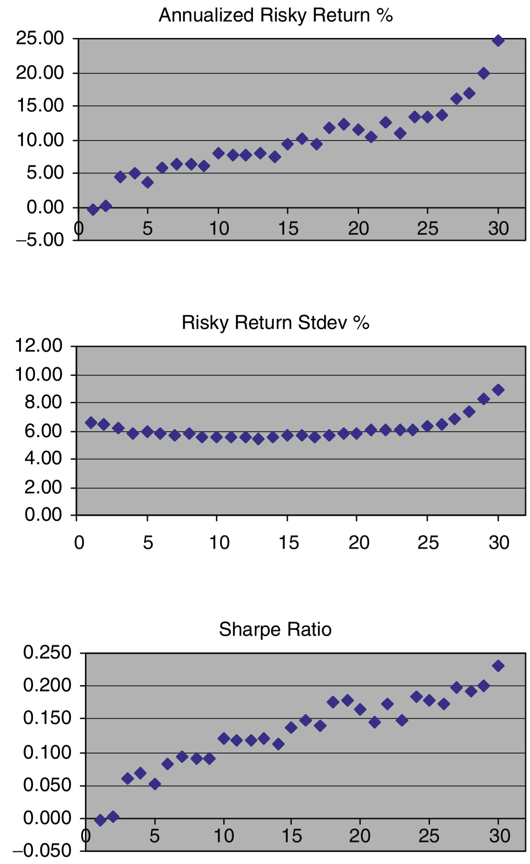
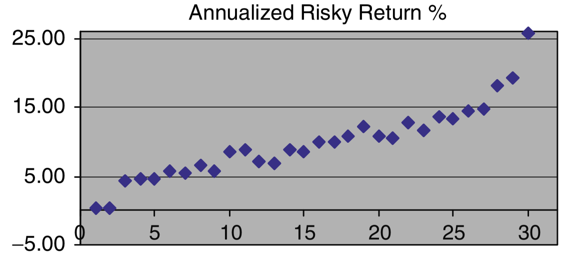
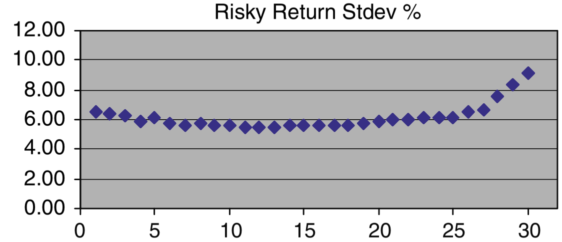


Exhibit 10.5K Annualized Monthly Average Returns Controls: Beta, Size, BP, FL, EP, DP, Sag5, Sug3, SI (Sales/Total Investment)

Portfolio	Forecast Score	Annualized Risky Return %	Risky Return Stdev %	Skewness	Sharpe Ratio
1	10.76	0.32	6.51	−0.449	0.004
2	11.73	0.33	6.39	−0.440	0.004
3	15.75	4.21	6.30	−0.497	0.056
4	20.74	4.67	5.90	−0.791	0.066
5	23.06	4.54	6.11	−0.475	0.062
6	25.98	5.64	5.75	−0.676	0.082
7	28.84	5.49	5.64	−0.674	0.081
8	31.26	6.43	5.71	−0.578	0.094
9	33.51	5.73	5.59	−0.533	0.085
10	36.01	8.61	5.62	−0.550	0.128
11	38.39	8.81	5.54	−0.647	0.132
12	40.70	7.09	5.51	−0.508	0.107
13	42.98	6.96	5.45	−0.633	0.106
14	45.51	8.69	5.56	−0.582	0.130
15	48.09	8.39	5.67	−0.514	0.123
16	50.44	9.91	5.65	−0.379	0.146
17	53.09	9.90	5.66	−0.233	0.146
18	55.79	10.84	5.60	−0.162	0.161
19	58.23	12.25	5.75	−0.240	0.177
20	61.01	10.92	5.82	−0.082	0.156
21	63.78	10.62	5.97	0.231	0.148
22	66.24	12.64	6.06	0.101	0.174
23	68.82	11.60	6.12	0.069	0.158
24	71.72	13.65	6.11	0.109	0.186
25	75.01	13.20	6.18	−0.032	0.178
26	78.45	14.36	6.57	0.474	0.182
27	81.04	14.72	6.71	0.338	0.183
28	85.49	18.03	7.59	0.774	0.198
29	89.58	19.30	8.30	1.360	0.194
30	92.80	25.75	9.08	1.657	0.236





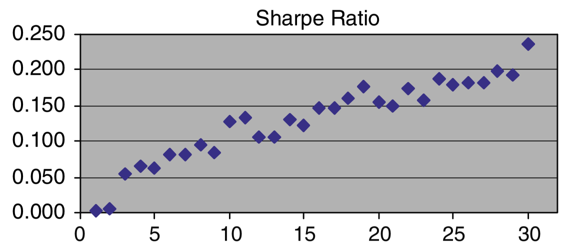
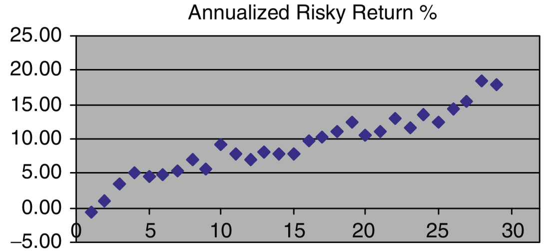
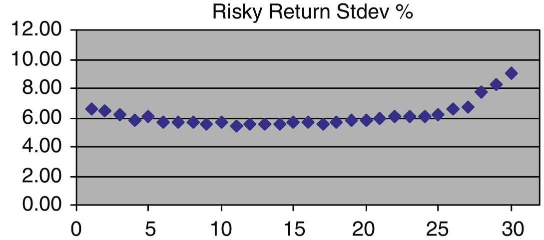


Exhibit 10.5L Annualized Monthly Average Returns Controls: Beta, Size, BP, FL, EP, DP, Sag5, Sug3, ROE, SI

Portfolio	Forecast Score	Annualized Risky Return %	Risky Return Stdev %	Skewness	Sharpe Ratio
1	10.89	−0.54	6.59	−0.503	−0.007
2	11.85	1.04	6.39	−0.444	0.014
3	15.86	3.34	6.21	−0.492	0.045
4	20.81	5.08	5.84	−0.718	0.072
5	23.12	4.63	6.01	−0.557	0.064
6	26.03	4.93	5.74	−0.667	0.072
7	28.89	5.40	5.72	−0.668	0.079
8	31.31	7.08	5.67	−0.620	0.104
9	33.54	5.56	5.61	−0.509	0.083
10	36.03	9.22	5.68	−0.593	0.135
11	38.41	7.95	5.41	−0.700	0.122
12	40.73	7.08	5.50	−0.521	0.107
13	43.01	8.00	5.50	−0.628	0.121
14	45.54	7.85	5.53	−0.590	0.118
15	48.11	7.92	5.69	−0.446	0.116
16	50.44	9.68	5.67	−0.268	0.142
17	53.09	10.18	5.58	−0.322	0.152
18	55.78	11.05	5.64	−0.347	0.163
19	58.24	12.33	5.79	−0.120	0.177
20	61.00	10.68	5.82	−0.044	0.153
21	63.76	11.14	5.98	0.059	0.155
22	66.22	13.11	6.07	0.261	0.180
23	68.79	11.53	6.07	0.014	0.158
24	71.68	13.52	6.11	0.165	0.184
25	74.97	12.48	6.16	−0.063	0.169
26	78.37	14.31	6.53	0.411	0.183
27	80.96	15.37	6.73	0.409	0.190
28	85.39	18.37	7.70	0.863	0.199
29	89.47	17.91	8.25	1.343	0.181
30	92.66	27.09	9.04	1.434	0.250





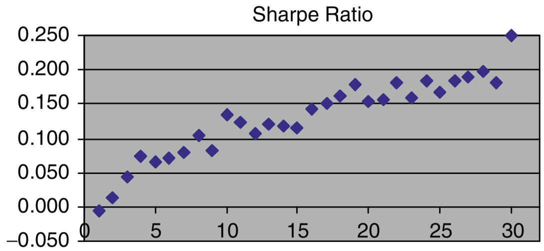


Exhibit 10.6 High-Minus-Low Values for Longrun Average Returns: 1968–2004

Control Variables	30-tiles	Deciles	Quintiles
No Constraints	0.088	0.074	0.065
β(Beta)	0.121	0.091	0.066
S(Size)	0.086	0.075	0.064
BP(Book-to-Market)	0.076	0.068	0.055
β, S, BP	0.087	0.079	0.056
FL, EP, DP	0.196	0.176	0.137
β, S, BP, FL, EP, DP	0.211	0.165	0.128
β, S, DP, FL, Sag5, Sug3	0.239	0.187	0.147
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE	0.267	0.193	0.143
β, S, BP, FL, EP, DP, Sag5, Sug3, ROI	0.250	0.190	0.143
β, S, BP, FL, EP, DP, Sag5, Sug3, SI	0.254	0.194	0.143
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE, SI	0.276	0.198	0.145
β, S, BP, FL, EP, DP, SAG5, SUG3, ROE, SI, CP, SP	0.191	0.142	0.111

Notes:
[1]·For each control set, the high-minus-low value for 30-tiles is computed by taking the difference in return between portfolio 30 and portfolio 1.
[2]·For deciles, the high-minus-low value is the difference between the average realized return for the top 3 portfolios minus the average for the bottom 3 portfolios.
[3]·For quintiles, the high-minus-low value is the difference between the average realized return in the top six portfolios minus the average realized return for the bottom six portfolios.

Sharpe ratios. The data tabulated in Exhibits 10.5A to 10.5L are the input to the tests in Exhibits 10.6 to 10.12.

2. *Exhibit 10.6* summarizes *high-minus-low return values*, the difference in long-run average return between the top and bottom control-matched fractile portfolios for each of the control constraints listed in Exhibit 10.4. In addition to the high-minus-low return differences for 30 tiles, Exhibit 10.6 also reports high-minus-low differences for deciles and quartiles. The rather large changes in high-minus-low differences for different control sets reflect cross-sectional interaction (correlation) between the return forecast and the different sets of control variables.
3. *Exhibits 10.7 to 10.11* summarize a series of cross-sectional regression fits that test how well the return forecast scores and/or return forecast rank explain the cross-section of realized average returns, the long-run realized Sharpe ratios for each fractile portfolio rank, the long-run realized standard deviations for each fractile portfolio rank, the long-run realized skewness coefficients for each fractile portfolio rank.
4. *Exhibit 10.12* is a stepwise regression assessment of how much realized standard deviation and realized skewness add to the ability of return forecast score to explain the cross-section of long-run realized returns. It is a test of alpha performance vs. return to risk (realized standard deviation) or skewness effects. To the extent that there are no material omissions in the full set of controls used in these regression tests, the difference in the significance of the regression coefficients indicates strongly that the very significant long-run return performance of

Exhibit 10.7 The Ability of Portfolio Number and Forecast Score to Explain the Cross Section of Realized Returns and Sharpe Ratios: 1968–2004

	$R_p = C_0 + C_1 p + \varepsilon_p$				$R_p = C_0 + C_1(FS) + \varepsilon_p$			
	slope	t	R^2	p-value	slope	t	R^2	p-value
No Constraints	.260	7.24	.652	<.0001	.079	7.23	.651	<.0001
β(Beta)	.272	7.42	.663	<.0001	.086	7.47	.666	<.0001
S(Size)	.263	7.91	.691	<.0001	.080	7.91	.691	<.0001
BP(Book-to-Market)	.219	7.63	.675	<.0001	.071	7.77	.683	<.0001
β, S, BP	.232	6.27	.584	<.0001	.079	6.44	.597	<.0001
FL, EP, DP	.538	17.6	.917	<.0001	.181	17.81	.919	<.0001
β, S, BP, FL, EP, DP	.523	16.82	.910	<.0001	.192	18.97	.928	<.0001
β, S, DP, FL, Sag5, Sug3	.602	15.44	.895	<.0001	.206	16.4	.906	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE	.578	12.99	.858	<.0001	.217	14.57	.884	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROI	.572	14.32	.880	<.0001	.215	16.44	.906	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, SI	.575	13.68	.870	<.0001	.216	15.66	.898	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE, SI	.584	12.7	.852	<.0001	.220	14.24	.879	<.0001
β, S,BP,FL,EP,DP,SAG5,SUG3,ROE,SI,CP,SP	.462	15.23	.892	<.0001	.186	16.56	.907	<.0001
	$100 \times (\text{Sharpe Ratio})_p = C_0 + C_1 p + \varepsilon_p$				$100 \times (\text{Sharpe Ratio})_p = C_0 + C_1(FS) + \varepsilon_p$			
	slope	t	R^2	p-value	slope	t	R^2	p-value
No Constraints	.078	6.97	.634	<.0001	.255	6.92	.631	<.0001
β(Beta)	.083	9.38	.759	<.0001	.265	9.37	.758	<.0001
S(Size)	.080	7.42	.663	<.0001	.262	7.36	.659	<.0001
BP(Book-to-Market)	.065	5.44	.514	<.0001	.201	5.46	.515	<.0001
β, S, BP	.072	7.48	.667	<.0001	.214	7.4	.662	<.0001
FL, EP, DP	.174	8.53	.722	<.0001	.530	9.26	.754	<.0001
β, S, BP, FL, EP, DP	.204	12.26	.843	<.0001	.564	12.87	.855	<.0001
β, S, DP, FL, Sag5, Sug3	.220	15.62	.897	<.0001	.650	16.56	.907	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE	.230	15	.889	<.0001	.621	14.86	.888	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROI	.229	15.82	.899	<.0001	.617	15.6	.897	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, SI	.227	16.86	.910	<.0001	.614	16.56	.907	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE, SI	.231	15.41	.895	<.0001	.624	15.18	.892	<.0001
β, S,BP,FL,EP,DP,SAG5,SUG3,ROE,SI,CP,SP	.212	11.29	.820	<.0001	.529	11.24	.819	<.0001

Exhibit 10.8 The Cross Sectional Dependence of Standard Deviation

$SD_p = C_0 + C_1 p + \varepsilon_p$	C_0	t	C_1	t	Adj-R^2	p-value
No Constraints	75.532	11.64	0.228	0.62	−0.022	0.538
β(Beta)	70.156	16.26	0.359	1.48	0.039	0.151
S(Size)	75.370	11.7	0.225	0.62	−0.022	0.541
BP(Book-to-Market)	73.528	11.42	0.316	0.87	−0.008	0.391
β, S, BP	68.963	16.9	0.396	1.72	0.063	0.096
FL, EP, DP	66.545	13.45	0.648	2.32	0.132	0.028
β, S, BP, FL, EP, DP	64.504	20.15	0.586	3.25	0.248	0.003
β, S, DP, FL, Sag5, Sug3	65.394	18.63	0.580	2.93	0.208	0.007
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE	64.420	19.95	0.593	3.26	0.249	0.003
β, S, BP, FL, EP, DP, Sag5, Sug3, ROI	64.634	20.82	0.579	3.31	0.256	0.003
β, S, BP, FL, EP, DP, Sag5, Sug3, SI	64.447	19.84	0.600	3.28	0.251	0.003
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE, SI	64.274	19.87	0.606	3.32	0.257	0.003
β, S,BP,FL,EP,DP,SAG5,SUG3,ROE,SI,CP,SP	66.078	29.13	0.421	3.3	0.254	0.003
$SD_p = C_0 + C_1 FS_p + \varepsilon_p$	C_0	t	C_1	t	Adj-R^2	p-value
No Constraints	75.741	11.97	0.067	0.61	−0.022	0.549
β(Beta)	70.105	16.21	0.113	1.49	0.040	0.149
S(Size)	75.545	11.96	0.067	0.6	−0.022	0.551
BP(Book-to-Market)	73.310	11.11	0.103	0.88	−0.008	0.385
β, S, BP	68.387	15.87	0.135	1.75	0.067	0.090
FL, EP, DP	64.662	12.48	0.238	2.59	0.164	0.015
β, S, BP, FL, EP, DP	62.149	17.51	0.228	3.56	0.288	0.001
β, S, DP, FL, Sag5, Sug3	63.970	17.26	0.208	3.15	0.236	0.004
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE	61.967	17.02	0.232	3.53	0.283	0.002
β, S, BP, FL, EP, DP, Sag5, Sug3, ROI	62.228	17.75	0.227	3.58	0.290	0.001
β, S, BP, FL, EP, DP, Sag5, Sug3, SI	61.974	16.96	0.235	3.55	0.286	0.001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE, SI	61.761	16.93	0.237	3.6	0.292	0.001
β, S,BP,FL,EP,DP,SAG5,SUG3,ROE,SI,CP,SP	63.872	23.11	0.174	3.44	0.272	0.002

Exhibit 10.9 The Cross Sectional Dependence of Standard Deviation: Linear and Quadratic Fits

$SD_p = C_0 + C_1 p + C_2(p-15)^2 + \varepsilon_p$	C_0	t	C_1	t	C_2	t	Adj-R^2	p-value
No Constraints	61.15	31.42	−0.01	−0.13	0.24	18.53	0.923	<.0001
β(Beta)	61.09	31.31	0.21	2.04	0.15	11.65	0.835	<.0001
S(Size)	61.09	31.54	−0.01	−0.14	0.24	18.48	0.922	<.0001
BP(Book-to-Market)	59.30	29.73	0.08	0.75	0.24	17.89	0.919	<.0001
β, S, BP	60.30	34.43	0.25	2.75	0.15	12.40	0.855	<.0001
FL, EP, DP	55.73	32.43	0.47	5.21	0.18	15.78	0.912	<.0001
β, S, BP, FL, EP, DP	57.74	41.1	0.47	6.46	0.11	12.08	0.878	<.0001
β, S, DP, FL, Sag5, Sug3	58.00	37.01	0.46	5.58	0.12	11.84	0.867	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE	57.68	38.08	0.48	6.08	0.11	11.15	0.861	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROI	58.09	42.16	0.47	6.54	0.11	11.92	0.877	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, SI	57.65	38.33	0.49	6.20	0.11	11.33	0.865	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE, SI	57.50	38.55	0.49	6.34	0.11	11.39	0.867	<.0001
β, S,BP,FL,EP,DP,SAG5,SUG3,ROE,SI,CP,SP	61.23	65.54	0.34	6.99	0.08	13.01	0.894	<.0001
$SD_p = C_0 + C_1(FS_p) + C_2(FS_p - mean(FS_p))^2 + \varepsilon_p$	C_0	t	C_1	t	C_2	t	Adj-R^2	p-value
No Constraints	57.45	28.27	0.07	2.30	0.02	18.21	0.920	<.0001
β(Beta)	58.89	30.66	0.11	3.74	0.01	12.45	0.852	<.0001
S(Size)	57.43	28.56	0.07	2.31	0.02	18.36	0.921	<.0001
BP(Book-to-Market)	56.29	32.94	0.11	3.89	0.02	22.24	0.946	<.0001
β, S, BP	58.00	37.39	0.14	5.50	0.02	15.46	0.902	<.0001
FL, EP, DP	54.72	33.76	0.18	6.46	0.02	17.32	0.928	<.0001
β, S, BP, FL, EP, DP	55.97	48.93	0.20	9.98	0.01	16.57	0.934	<.0001
β, S, DP, FL, Sag5, Sug3	56.60	38.15	0.18	7.18	0.01	13.27	0.895	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE	55.46	42.5	0.21	9.24	0.01	14.80	0.918	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROI	55.92	48.3	0.20	10.20	0.01	16.20	0.931	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, SI	55.42	42.93	0.21	9.44	0.01	15.04	0.921	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE, SI	55.22	43.02	0.21	9.62	0.01	15.11	0.922	<.0001
β, S,BP,FL,EP,DP,SAG5,SUG3,ROE,SI,CP,SP	58.80	67.67	0.17	11.06	0.01	17.13	0.936	<.0001

Exhibit 10.10 The Cross Sectional Dependence of Realized Portfolio Skewness

$S_p = C_0 + C_1 p + \varepsilon_p$	C_0	t	C_1	t	Adj-R^2	p-value
No Constraints	−0.702	−3.22	0.047	3.86	0.323	0.0006
β(Beta)	−0.890	−5.99	0.054	6.4	0.580	<.0001
S(Size)	−0.697	−3.16	0.047	3.79	0.315	0.0007
BP(Book-to-Market)	−0.782	−3.52	0.050	3.98	0.338	0.0004
β, S, BP	−0.945	−7.03	0.054	7.13	0.632	<.0001
FL, EP, DP	−0.652	−2.87	0.043	3.34	0.259	0.0024
β, S, BP, FL, EP, DP	−0.993	−8.12	0.054	7.78	0.672	<.0001
β, S, DP, FL, Sag5, Sug3	−0.975	−7.26	0.056	7.35	0.647	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE	−1.009	−8.36	0.055	8.06	0.688	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROI	−1.015	−9.08	0.055	8.74	0.722	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, SI	−1.005	−7.89	0.056	7.77	0.672	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE, SI	−1.008	−8.5	0.055	8.28	0.700	<.0001
β, S,BP,FL,EP,DP,SAG5,SUG3,ROE,SI,CP,SP	−0.836	−11.55	0.040	9.8	0.766	<.0001
$S_p = C_0 + C_1 FS_p + \varepsilon_p$	C_0	t	C_1	t	Adj-R^2	p-value
No Constraints	−0.674	−3.17	0.014	3.83	0.321	0.0007
β(Beta)	−0.895	−6.02	0.017	6.43	0.582	<.0001
S(Size)	−0.676	−3.12	0.014	3.77	0.313	0.0008
BP(Book-to-Market)	−0.804	−3.52	0.016	3.96	0.336	0.0005
β, S, BP	−1.007	−7.08	0.018	7.12	0.632	<.0001
FL, EP, DP	−0.751	−3.17	0.015	3.62	0.295	0.0012
β, S, BP, FL, EP, DP	−1.163	−8.75	0.020	8.34	0.703	<.0001
β, S, DP, FL, Sag5, Sug3	−1.080	−7.83	0.019	7.86	0.677	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE	−1.196	−9	0.021	8.6	0.716	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROI	−1.205	−9.84	0.021	9.4	0.751	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, SI	−1.196	−8.55	0.021	8.33	0.702	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE, SI	−1.198	−9.19	0.021	8.86	0.728	<.0001
β, S,BP,FL,EP,DP,SAG5,SUG3,ROE,SI,CP,SP	−1.023	−11.56	0.016	9.92	0.770	<.0001

Exhibit 10.11 The Cross Sectional Dependence of Realized Portfolio Skewness: Linear and Quadratic Fits

$S_p = C_0 + C_1 p + C_2(p - 15)^2 + \boldsymbol{\varepsilon}_p$	C_0	t	C_1	t	C_2	t	Adj-R^2	p-value
No Constraints	−1.176	−14.81	0.039	9.53	0.008	14.98	0.925	<.0001
β(Beta)	−1.178	−13.14	0.049	10.46	0.005	8.04	0.872	<.0001
S(Size)	−1.172	−13.7	0.039	8.76	0.008	13.93	0.913	<.0001
BP(Book-to-Market)	−1.260	−14.27	0.042	9.1	0.008	13.58	0.912	<.0001
β, S, BP	−1.220	−17.88	0.049	13.9	0.005	10.11	0.920	<.0001
FL, EP, DP	−1.115	−9.51	0.035	5.73	0.008	9.9	0.834	<.0001
β, S, BP, FL, EP, DP	−1.222	−15.54	0.050	12.14	0.004	7.31	0.886	<.0001
β, S, DP, FL, Sag5, Sug3	−1.250	−18.23	0.051	14.29	0.005	10.04	0.923	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE	−1.235	−15.94	0.051	12.64	0.004	7.33	0.892	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROI	−1.236	−19.21	0.051	15.34	0.004	8.58	0.923	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, SI	−1.250	−16.09	0.052	12.76	0.004	7.91	0.897	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE, SI	−1.243	−18.37	0.051	14.59	0.004	8.68	0.918	<.0001
β, S,BP,FL,EP,DP,SAG5,SUG3,ROE,SI,CP,SP	−0.960	−17.78	0.038	13.48	0.002	5.73	0.891	<.0001
$S_p = C_0 + C_1(FS_p) + C_2(FS_p - \mathit{mean}(FS_p))^2 + \boldsymbol{\varepsilon}_p$	C_0	t	C_1	t	C_2	t	Adj-R^2	p-value
No Constraints	−1.272	−14.36	0.014	10.7	0.001	13.67	0.911	<.0001
β(Beta)	−1.252	−14.06	0.017	12.13	0.000	8.55	0.883	<.0001
S(Size)	−1.274	−13.42	0.014	9.95	0.001	12.84	0.900	<.0001
BP(Book-to-Market)	−1.375	−15.84	0.016	11.7	0.001	14.68	0.923	<.0001
β, S, BP	−1.340	−21.18	0.018	17.84	0.000	12.14	0.941	<.0001
FL, EP, DP	−1.164	−9.15	0.013	5.92	0.001	9.16	0.822	<.0001
β, S, BP, FL, EP, DP	−1.365	−17.42	0.019	14.07	0.000	7.93	0.907	<.0001
β, S, DP, FL, Sag5, Sug3	−1.347	−21.04	0.018	17.12	0.001	11.15	0.940	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE	−1.408	−18.3	0.020	15.07	0.000	8.2	0.916	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROI	−1.411	−23.11	0.020	19.17	0.000	10	0.945	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, SI	−1.427	−19.23	0.020	15.88	0.001	9.24	0.926	<.0001
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE, SI	−1.418	−21.82	0.020	18.03	0.000	10.02	0.940	<.0001
β, S,BP,FL,EP,DP,SAG5,SUG3,ROE,SI,CP,SP	−1.150	−18.14	0.016	14.51	0.000	5.89	0.896	<.0001

Exhibit 10.12 Stepwise Regression Summaries: The Cross Sectional Dependence Average Portfolio Returns On Forecast Score, Standard Deviation, and Skewness

$R_p = C_0 + C_1(FS_p) + \boldsymbol{\varepsilon}_p$

	C_0	t	C_1	t	Adj-R^2
No Constraints	6.220	9.99	0.079	7.23	0.64
β (Beta)	5.762	8.86	0.086	7.47	0.65
S(Size)	6.146	10.62	0.080	7.91	0.68
BP(Book-to-Market)	6.669	12.91	0.071	7.77	0.67
β, S, BP	6.237	9.13	0.079	6.44	0.58
FL, EP, DP	0.610	1.06	0.181	17.81	0.92
β, S, BP, FL, EP, DP	0.016	0.03	0.192	18.97	0.52
β, S, DP, FL, Sag5, Sug3	−0.571	−0.81	0.206	16.4	0.49
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE	−1.121	−1.36	0.217	14.57	0.88
β, S, BP, FL, EP, DP, Sag5, Sug3, ROI	−1.005	−1.39	0.215	16.44	0.90
β, S, BP, FL, EP, DP, Sag5, Sug3, SI	−1.007	−1.32	0.216	15.66	0.89
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE, SI	−1.218	−1.43	0.220	14.24	0.94
β, S,BP,FL,EP,DP,SAG5,SUG3,ROE,SI,CP,SP	0.317	0.51	0.186	16.56	0.90

$R_p = C_0 + C_1(FS_p) + C_2(SD_p - \mathrm{mean}(SD_p) + C_3(SD_p - \mathit{mean}(SD_p))^2 + \boldsymbol{\varepsilon}_p$

	C_0	t	C_1	t	C_2	t	C_3	t	Adj-R^2
No Constraints	6.301	14.74	0.071	9.63	0.052	2.54	0.001	1.38	0.84
β (Beta)	6.324	14.92	0.069	8.81	0.076	2.38	0.002	1.55	0.86
S(Size)	6.243	16.4	0.073	11.07	0.052	2.88	0.001	1.32	0.87
BP(Book-to-Market)	6.829	22.48	0.061	11.11	0.035	2.48	0.001	2.48	0.89
β, S, BP	7.036	14.81	0.061	6.87	0.112	3.24	0.000	0.34	0.82
FL, EP, DP	−0.114	−0.32	0.178	28.18	−0.114	−6.6	0.005	8.05	0.97
β, S, BP, FL, EP, DP	−0.075	−0.14	0.182	18.69	−0.093	−2.35	0.006	4.01	0.95
β, S, DP, FL, Sag5, Sug3	−0.263	−0.54	0.185	21.13	−0.065	−1.85	0.007	5.77	0.97
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE	−0.483	−0.71	0.188	15.41	−0.061	−1.22	0.009	4.32	0.94
β, S, BP, FL, EP, DP, Sag5, Sug3, ROI	−0.509	−0.93	0.190	19.1	−0.067	−1.65	0.009	5.23	0.96
β, S, BP, FL, EP, DP, Sag5, Sug3, SI	−0.244	−0.48	0.185	20.01	−0.051	−1.36	0.008	5.73	0.97
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE, SI	−0.562	−0.91	0.188	16.91	−0.079	−1.75	0.010	5.53	0.96
β, S,BP,FL,EP,DP,SAG5,SUG3,ROE,SI,CP,SP	−0.219	−0.34	0.188	15.71	−0.145	−2.58	0.010	3.05	0.92

(continued)

Exhibit 10.12 (continued)

$R_p = C_0 + C_1(FS_p) + C_2(SD_p - \text{mean}(SD_p) + C_3(SD_p - \mathit{mean}(SD_p))^2 + C_4(S_p - \text{mean}(S_p) + \boldsymbol{\varepsilon}_p$

	C_0	t	C_1	t	C_2	t	C_3	t	C_4	t	Adj-R^2
No Constraints	7.42	8.99	.048	2.94	−.011	−0.24	.001	1.53	1.88	1.57	0.85
β(Beta)	4.80	5.2	.101	5.34	.158	2.92	.002	1.29	−2.34	−1.84	0.88
S(Size)	7.81	12.53	.041	3.35	−.035	−1.06	.001	1.65	2.64	2.98	0.90
BP(Book-to-Market)	7.61	12.98	.045	3.79	−.010	−0.3	.001	2.72	1.26	1.54	0.90
β, S, BP	5.67	3.81	.090	2.91	.184	2.23	.000	0.01	−2.01	−0.97	0.82
FL, EP, DP	0.12	0.31	.173	24.08	−.157	−4.45	.005	8.3	0.97	1.39	0.98
β, S, BP, FL, EP, DP	0.14	0.13	.178	8.2	−.104	−1.66	.006	3.82	0.35	0.22	0.95
β, S, DP, FL, Sag5, Sug3	−0.65	−0.53	.193	7.73	−.038	−0.45	.007	5.32	−0.67	−0.34	0.96
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE	−1.90	−1.28	.216	7.4	.015	0.17	.009	4.29	−2.21	−1.07	0.94
β, S, BP, FL, EP, DP, Sag5, Sug3, ROI	0.41	0.26	.172	5.59	−.113	−1.36	.009	5.16	1.38	0.63	0.96
β, S, BP, FL, EP, DP, Sag5, Sug3, SI	−1.51	−1.26	.210	8.94	.021	0.29	.008	5.77	−2.00	−1.16	0.97
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE, SI	−0.57	−0.36	.188	5.85	−.079	−0.79	.010	5.21	−0.01	0	0.95
β, S,BP,FL,EP,DP,SAG5,SUG3,ROE,SI,CP,SP	0.78	0.59	.168	6.55	−.189	−2.5	.010	3.04	1.71	0.88	0.92

$R_p = C_0 + C_1(FS_p) + C_2(SD_p - \text{mean}(SD_p) + C_3(SD_p - \mathit{mean}(SD_p))^2 + C_4(S_p - \text{mean}(S_p) + C_5(S_p - \text{mean}(S_p))^3 + \boldsymbol{\varepsilon}_p$

	C_0	t	C_1	t	C_2	t	C_3	t	C_4	t	C_5	t	Adj-R^2
No Constraints	7.41	9.27	.053	3.29	.021	0.44	.000	−0.28	1.04	0.81	0.34	1.61	0.86
β (Beta)	4.80	5.21	.101	5.35	.157	2.92	.000	−0.06	−2.91	−2.12	0.74	1.08	0.88
S(Size)	7.80	12.85	.045	3.68	−.012	−0.33	.000	−0.13	2.00	2.09	0.25	1.55	0.91
BP(Book-to-Market)	7.70	12.77	.044	3.69	−.006	−0.18	.001	1.13	1.13	1.33	0.14	0.77	0.90
β, S, BP	6.03	4.34	.089	3.11	.201	2.61	−.008	−2.09	−3.17	−1.59	2.57	2.25	0.84
FL, EP, DP	0.17	0.43	.170	20.67	−.169	−4.24	.006	3.6	1.33	1.48	−0.23	−0.66	0.97
β, S, BP, FL, EP, DP	−0.51	−0.36	.193	6.73	−.065	−0.81	.004	1.37	−0.85	−0.39	0.40	0.79	0.95
β, S, DP, FL, Sag5, Sug3	0.71	0.48	.167	5.63	−.097	−1.06	.010	4.28	2.06	0.8	−1.15	−1.54	0.97
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE	1.19	0.7	.153	4.47	−.126	−1.38	.015	5.23	3.06	1.18	−1.79	−2.85	0.96
β, S, BP, FL, EP, DP, Sag5, Sug3, ROI	0.14	0.08	.179	5.23	−.098	−1.09	.006	1.35	0.82	0.33	0.51	0.48	0.96
β, S, BP, FL, EP, DP, Sag5, Sug3, SI	−1.40	−1.08	.207	7.75	.016	0.21	.010	1.58	−1.74	−0.84	−0.31	−0.23	0.97
β, S, BP, FL, EP, DP, Sag5, Sug3, ROE, SI	0.98	0.65	.154	4.98	−.131	−1.44	.019	5.03	3.25	1.37	−3.09	−2.76	0.96
β, S,BP,FL,EP,DP,SAG5,SUG3,ROE,SI,CP,SP	0.09	0.06	.184	5.48	−.146	−1.54	.004	0.47	0.10	0.04	2.25	0.76	0.92

our test of the Bloch et al. (1993) stock return forecast model is primarily alpha performance and not a return to greater risk bearing to the extent that long-run cross-time realized standard deviation is an appropriate measure of risk bearing for the 456 months in 1967–2004.

5. *Exhibits 10.13 to 10.15* compare realized Sharpe ratios to index returns for an upper quartile portfolio (top six fractile portfolios averaged together) for three variations in control variables and four subperiods as well as the overall time period. These results support the conclusion stated in #4 above of alpha performance potential for the Bloch et al. (1993) model, especially given the positive skewness of the upper quartile portfolio relative to the skewness of the sample.

10.7.4 Cross Sections: Overview

An informative sequence of control restrictions logically begins with no constraints. Exhibit 10.5A summarizes the return-forecast cross sections for the case of no control constraints. Exhibits 10.5B to 10.5O summarize the return-forecast cross sections for the control sets in the order listed in Exhibit 10.4. The various cross-sectional plots give an overview of how the apparent dependency on the return forecast changes with changes in the control sets.

Before reviewing the different combinations of control constraints, it is useful to provide a summary overview of the cross-sectional dependencies. The long-run average realized return cross sections all show that the realized returns rank order well with return forecast with no controls and for each of the control sets. However, as already noted in our overview in Sect. 10.2, the range of realized returns (as measured by high-minus-low return values) shows dramatic change with changes in the control sets.

The cross-section of realized standard deviations is remarkably smooth on a portfolio-to-portfolio basis, especially after imposing both risk and tax controls. The realized standard deviations are higher for both the extreme low forecast portfolios and the extreme high forecast portfolios. The overall dependency is nonlinear and non-monotonic. The plots suggest a quadratic dependency on forecasted return rank or score. The range of realized standard deviations is relatively small, less than 4% once risk and tax controls are imposed. This relatively small range for standard deviations relative to the much larger range for the cross-section of realized returns suggests that the cross-section of Sharpe ratios should be upward sloping with significant cross-sectional range. The plots of Sharpe ratios confirm this conjecture. The plots of Sharpe ratios show more scatter than either the returns or standard deviations but do show an overall increase with forecast score. Once both risk and tax controls are imposed, the range in Sharpe ratios goes from a low near zero to more than 0.20.

Exhibit 10.13 SHARPE RATIOS: No Control Constraints

Overall Time Period: January 1967 to December 2004

CRSP Value-Weighted Index			**CRSP Equal-Weighted Index**			**Lower Quintile Portfolio**			**Upper Quintile Portfolio**		
Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio
0.48	4.60	0.105	0.81	5.91	0.138	0.69	6.83	0.101	1.17	7.56	0.155

Sub-Period 1: January 1967 to December 1974

CRSP Value-Weighted Index			**CRSP Equal-Weighted Index**			**Lower Quintile Portfolio**			**Upper Quintile Portfolio**		
Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio
−0.13	5.02	−0.027	0.11	7.42	0.015	−0.29	6.85	−0.042	−0.20	7.62	−0.026

Sub-Period 2: January 1975 to December 1984

CRSP Value-Weighted Index			**CRSP Equal-Weighted Index**			**Lower Quintile Portfolio**			**Upper Quintile Portfolio**		
Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio
0.59	4.46	0.131	1.38	5.58	0.248	1.09	6.98	0.157	1.95	6.67	0.293

Sub-Period 3: January 1985 to December 1994

CRSP Value-Weighted Index			**CRSP Equal-Weighted Index**			**Lower Quintile Portfolio**			**Upper Quintile Portfolio**		
Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio
0.66	4.34	0.153	0.49	4.80	0.102	0.52	5.99	0.087	1.03	5.73	0.181

Sub-Period 4: January 1995 to December 2004

CRSP Value-Weighted Index			**CRSP Equal-Weighted Index**			**Lower Quintile Portfolio**			**Upper Quintile Portfolio**		
Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio
0.69	4.65	0.148	1.13	5.87	0.193	1.25	7.42	0.168	1.64	9.61	0.170

Exhibit 10.14 SHARPE RATIOS Constraints: Beta, Size, DP, FL, SAG5, SUG3

Overall Time Period: January 1967 to December 2004

CRSP Value-Weighted Index			**CRSP Equal-Weighted Index**			**Lower Quintile Portfolio**			**Upper Quintile Portfolio**		
Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio
0.48	4.60	0.105	0.81	5.91	0.138	0.25	5.85	0.043	1.40	6.95	0.201

Sub-Period 1: January 1967 to December 1974

CRSP Value-Weighted Index			**CRSP Equal-Weighted Index**			**Lower Quintile Portfolio**			**Upper Quintile Portfolio**		
Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio
−0.13	5.02	−0.027	0.11	7.42	0.015	−1.44	5.99	−0.240	0.55	7.21	0.076

Sub-Period 2: January 1975 to December 1984

CRSP Value-Weighted Index			**CRSP Equal-Weighted Index**			**Lower Quintile Portfolio**			**Upper Quintile Portfolio**		
Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio
0.59	4.46	0.131	1.38	5.58	0.248	0.36	5.94	0.061	2.19	6.55	0.335

Sub-Period 3: January 1985 to December 1994

CRSP Value-Weighted Index			**CRSP Equal-Weighted Index**			**Lower Quintile Portfolio**			**Upper Quintile Portfolio**		
Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio
0.66	4.34	0.153	0.49	4.80	0.102	0.58	5.48	0.106	1.01	5.59	0.181

Sub-Period 4: January 1995 to December 2004

CRSP Value-Weighted Index			**CRSP Equal-Weighted Index**			**Lower Quintile Portfolio**			**Upper Quintile Portfolio**		
Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio
0.69	4.65	0.148	1.13	5.87	0.193	1.16	5.78	0.201	1.66	8.22	0.202

Exhibit 10.15 SHARPE RATIOS Constraints: β, S, BP, FL, EP, DP, Sag5, Sug3, ROE, SI

Overall Time Period: January 1967 to December 2004

CRSP Value-Weighted Index			**CRSP Equal-Weighted Index**			**Lower Quintile Portfolio**			**Upper Quintile Portfolio**		
Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio
0.48	4.60	0.105	0.81	5.91	0.138	0.32	5.76	0.055	1.39	6.83	0.204

Sub-Period 1: January 1967 to December 1974

CRSP Value-Weighted Index			**CRSP Equal-Weighted Index**			**Lower Quintile Portfolio**			**Upper Quintile Portfolio**		
Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio
−0.13	5.02	−0.027	0.11	7.42	0.015	−1.27	5.88	−0.216	0.42	7.20	0.058

Sub-Period 2: January 1975 to December 1984

CRSP Value-Weighted Index			**CRSP Equal-Weighted Index**			**Lower Quintile Portfolio**			**Upper Quintile Portfolio**		
Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio
0.59	4.46	0.131	1.38	5.58	0.248	0.44	5.88	0.075	2.21	6.53	0.338

Sub-Period 3: January 1985 to December 1994

CRSP Value-Weighted Index			**CRSP Equal-Weighted Index**			**Lower Quintile Portfolio**			**Upper Quintile Portfolio**		
Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio
0.66	4.34	0.153	0.49	4.80	0.102	0.61	5.40	0.113	1.04	5.47	0.190

Sub-Period 4: January 1995 to December 2004

CRSP Value-Weighted Index			**CRSP Equal-Weighted Index**			**Lower Quintile Portfolio**			**Upper Quintile Portfolio**		
Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio	Avg Rtn%	Std Dev	Sharpe Ratio
0.69	4.65	0.148	1.13	5.87	0.193	1.17	5.71	0.204	1.70	7.92	0.215

10.7.5 How Cross Sections Change with Changes in the Control Sets

On the assumption that systematic risk should be separated from excess returns, we logically move from no controls to imposing just risk controls. Given their importance, we assess the effect of systematic risk controls by just adding a single control for standard risk variables (beta only, size only, and book-price only). Then we logically control for combinations of risk variables, i.e., the Fama–French three risk factors together: beta, size, and book-to-price. Exhibits 10.5B to 10.5E summarize the return-forecast cross sections for the cases of just a single systematic risk control and the three Fama–French risk controls together.

We call readers attention to the fact the Fama–French model includes a variable from the return forecast model, namely BP, the current book-price ratio (book-to-market ratio). Hence, this control removes any forecast performance benefit and any risk effect related to the book-to-market ratio. What remains in this cross section summarized in Exhibit 10.5E is the performance value of other seven model variables while controlling for beta, size, and the BP ratio.

After assessing systematic risk controls alone and in combination, we turn next to the tax controls. Recall that the three tax controls are DP, EP, and FL. We impose the three tax controls together but without any systematic risk controls in Exhibit 10.5F. The long-run return cross sections with just the three tax controls are summarized in Exhibit 10.5F.

After assessing risk controls and tax controls individually, we next combine the three systematic risk controls and the three tax controls. These cross sections are summarized in Exhibit 10.5G. This set of six risk-tax controls should produce a realized return cross section with realized returns isolated from our a priori assumption of systematically priced risk and tax effects. To the extent that we have correctly identified systematic risk factor controls and tax factor controls,[21] the rest of the controls should reflect primarily firm-specific return effects.

We call readers attention to the fact that the tax controls include a variable from the return forecast model, namely the current earnings yield **EP**, which some also view as a risk instrument complementary to the book-price ratio. Hence, imposing the three tax controls as summarized in Exhibit 10.5F and the combination of risk and tax controls as summarized in Exhibit 10.5G means that any forecast performance benefit related to the earnings yield is removed from these realized return cross sections. When the six factor risk-tax controls are imposed (control set F), any performance from both **BP** and **EP** are removed from the realized return cross section and what remains in the realized return cross section is the performance effect of just the remaining six of the eight forecast model variables.

We can assess the impact of suppressing any contribution of these two forecast model controls on the cross-section of realized returns by looking at a risk-tax

[21] The six risk-tax controls used here are very similar to the risk controls in the recently published BARRA Global Equity Model GEM2 as reported in Menchero et al. (2009) in this volume.

controlled cross section without these two forecast model controls. Control set G (summarized in Exhibit 10.5G) is the full intersection of the tax controls (FL, EP, DP) and the Fama–French risk controls.

10.7.6 More Cross Sections: Growth and Other Firm-Specific Controls

After imposing both the risk and the tax controls to remove systematic risk and tax effects from the observed cross sections, the obvious next candidates for firm-specific control is growth.[22] Controlling for expectations of future growth is clearly important to have a well-isolated return-forecast cross section given: (1) the conventional view that there is a value-growth trade-off, and (2) a significant negative sample correlation between our value variables and our two growth measures in many of the sample months.

Exhibit 10.2 defines two firm-specific growth controls: 5-year past sales growth and 3-year past sustainable growth. There is not a consensus on how to measure growth for explaining returns, for predicting future returns, or especially for our problem of using growth controls to remove the impact of growth expectations/realizations from our return-forecast cross sections. Given that there is not a consensus on how best to measure and control for firm-specific growth effects, our choice of our two growth controls merits explanation.

There are several reasons for using two growth controls. First, as in Lakonishok et al. (1994), we proxy future growth by a rolling value of past sales growth from the most recently reported past 5 years sales. However, as reported in Lakonishok et al. (1994), this growth proxy seems to be a relatively poor predictor of future growth. Our own tests of sales growth proxies indicate that 5-year sales growth is a better predictor of future earnings growth and cashflow growth than are either past earnings growth itself or past cashflow growth itself. However, compared with the ex ante portfolio-level predictive value of other control variables such as beta, size, book-price, dividend-price, and even financial structure, our assessment of future predictive value of sales growth concurred with the negative assessment of Lakonishok et al. (1994) of relatively poor predictive value for future growth in sales, earnings, or cashflow, especially for the top two growth deciles. Our analysis suggests that at least part of the explanation of poor predictive value (beyond the fact that future growth is generally recognized to be hard to predict) arises when the past sales growth rate is well above the rate that can be sustained from internally generated funds.

[22] We have referred to growth as a firm-specific variable. There is some controversy over whether growth is a performance measure or whether it should be viewed as a risk variable. Whether growth is performance or risk or a combination does not matter in terms of using it as a control to isolate return-forecast dependencies from growth effects regardless of whether growth is performance, risk, or both. Once we impose growth controls, we remove both the performance and the risk impact on the cross sections.

Given the limitations of past sales growth, there appears to be a need for greater accuracy in predictors of future growth expectations. Rather than discarding past sales growth, we decided to add a second growth control to improve the ability to control for crosssectional differences in future growth and especially to reflect the breakdown in past sales growth rates when firms have been growing at rates above the sustainable rate. For our second growth control, we decided to focus on the idea of sustainable growth.[23] A company's *sustainable growth rate* is defined as the growth rate that a firm could sustain from internally generated funds assuming no change in financial structure or profitability and also assuming proportional growth in all assets. However, rather than the one-period textbook measure of sustainable growth, we have followed the standard analyst practice of using a longer time period to obtain a smoothed, normalized measure of sustainable growth. We used the past *3-year sustainable growth rate* defined as the past 3-years of *additions to retained earnings* divided by the most recent (nearest 2-month back annual financial statement value of *total common equity* (book value)). Thus, for our growth controls we add to the rolling 5-year back sales growth rate a second growth control, a rolling 3-year back sustainable growth rate (3-year average addition to retained earnings per current dollar of total common equity).

We add our two growth controls to other variable combinations and especially to our risk-tax controls. Control set H (summarized in Exhibit 10.5H) is control set G (six factor risk-tax controls together) less the two value controls BP and EP plus the two growth controls alone.

Exhibits 10.7J, 10.7K, and 10.7L each add one additional control to the six risk-tax controls and two growth controls. These constraint sets each add one additional firm-specific control variable to reflect profitability and/or capital intensity controls. They can be viewed as additional controls on growth quality and/or forecast quality itself. These controls have some effect – larger high-minus-low values, less low-end standard deviation, and somewhat greater skewness. The fact that adding these additional control variables to the set of risk, tax, and growth controls seems to have relatively low impact (compared with the earlier control variables) on the cross sections summarized in Exhibits 10.7J through 10.7L is an indication that the prior controls are relatively complete in terms of isolating the return forecast from other firm-specific return impact variables.

In Exhibits 10.H to 10.L, we see that the two growth controls increase return range by about 5%, mainly by increasing realized returns for the portfolios with high forecasted returns. Clearly growth controls impact the value-focused cross-section. Growth and value are correlated.

[23] The merit of using sustainable growth as an additional growth control is reflecting the fact that companies growing faster than the sustainable growth rate in the past 5 years must either slow down, dilute existing shareholders by issuing equity to maintain the current debt-equity mix, or increase use of debt financing with an associated increase in overall systematic risk and especially downside risk at high debt levels, or possibly a combination of all three of these return-negative actions.

The final constraint set in the cross sections summarized in Exhibits 10.5A to 10.5L is all of the control variables except for ROI.[24] Interestingly, this most complete control set has the highest high-minus-low range, more than 27%. However, in the plots of realized return vs. forecast score, this top return of 27.09% for portfolio #30 is not plotted because it exceeds the graph range of 25%.

10.8 High-Minus-Low Differences

Exhibit 10.6 summarizes high-minus-low returns for major constraint sets. The first column names the constraint set. The next three columns give high-minus-low returns. For 30 fractile portfolios, this value is the long-run average return on portfolio 30 (the fractile portfolio with the highest value forecast score) less the long-run average return on portfolio 1 (the fractile with the lowest value forecast score). The next column is the average return for the two highest fractiles minus the average return for the two lowest fractiles. The next column is the average of the top three minus the average of the bottom three fractiles. With 30 fractile portfolios in the cross sections of conditional return response observations, the difference for the top three returns combined and the bottom three portfolios combined represents a high-minus-low return for the top and bottom deciles of the cross-section.

Since all portfolios in each cross-section are matched to the ex ante values of the listed factor controls, the high-minus-low values are the long-run realized returns on a factor neutral arbitrage portfolio, i.e., a portfolio that is long in one or more of the top 30 fractile portfolios and short in the corresponding low score portfolios. It is "*factor neutral*" in the ex ante values of each of the imposed control variables because each of the portfolios in each cross-section has been matched to the sample average value of the imposed controls. Therefore, a long-short combination of any two portfolios has zero ex ante exposure to the imposed controls.

The high-minus-low values are the annualized average of 458 monthly values. Thus, they indicate the economic significance of the composite value score before any transaction costs for a naive factor-neutral portfolio strategy.

The team "*naïve*" refers to the fact that these portfolios are formed on the basis of return data alone without using any information about variances, covariances, or higher moments such as skewness. Given that past values of both variance and co-variance are fairly good predictions of month-ahead values, use of the value-focused return scores with mean–variance optimization should always produce superior market-neutral hedge portfolios in terms of Sharpe ratios. The Sharpe ratios reported

[24] The technical reason for excluding ROI is that controlling for both ROE and ROI simultaneously can produce near singular matrices in the optimization program. The economic reason is that ROI is an overall firm return for all providers of financing while ROE seems to be a more direct period-to-period accounting measure of common equity return.

for these three market neutral portfolios are lower bounds on the mean–variance optimized factor neutral portfolios.

The high-minus-low ranges and associated Sharpe ratios both exhibit a strong dependency on the imposed controls. The unranked portfolio has a range of 8%, the Fama–French three-factor control set has a range of 9.6% while just imposing the three "tax controls" has a range of 12.7%. Most dramatic is the six-factor control set of Fama–French plus tax controls. The hml range increases to almost 19%.

Adding growth controls increases the hml to more than 24%, triple the range for the unranked cross-section and more than double the range when the only control variables are the conventional Fama–French three-factor risk instruments.

Adding sales intensity and profitability controls further increases the range.

The three tax controls are especially important in increasing range.

10.8.1 Summary Comments

From these results, we conclude that ranking ordering into fractiles does not average out factor risk in the extremes of the value fractiles. Likewise, the Fama–French three-factor risk controls actually control far less than half of the removable factor variation. In the results reported here, tax controls are more important than risk controls.

The plots of standard deviation show higher values for each tail of the cross-section. This pattern of ex post standard deviation, first declining slightly and then being relatively flat with a rise at the right end is not the near monotonic cross-sectional pattern that we would expect to see if the return forecast score was a risk instrument. These exhibits tabulate the skewness coefficient and show an increasing right skewness as the forecast score increases. The combination of a non-monotonic standard deviation and increasing right skewness means that the issue of risk and value is complex and clearly needs more systematic research.

We remind readers that the case of *all controls* (Exhibit 10.5O) means suppressing cross-sectional variation in both the ex ante book-to-market ratio and the ex ante earnings yield by making both of these variables into controls. Removing both the book-to-market and earnings price effect does not significantly reduce the return range or the Sharpe ratios. Consistent with the value benefit assertions of most value managers, the turning-point-focused value measures, current relative to 5-year average in the forecast model, seem to be the primary source of both the large high-minus-low return range and the superior Sharpe ratios.

10.9 Regression Tests

Exhibits 10.7 through 10.11 summarize regression tests of the ability of the return forecast score (return forecast rank) to explain the long run realized cross-sections of average returns, Sharpe ratios, standard deviations, and skewness coefficients.

10.9.1 The Cross-Section of Realized Returns and Sharpe Ratios

For each set of control constraints, Exhibit 10.7 summarizes regressions of the long-run 456 month average value of realized risky return on the average value of return forecast score and also on the portfolio relative-rank-order number. The results in almost every case are very similar whether we use average return forecast score or rank-order number as the explanatory variable.

All of the regressions in Exhibit 10.7 have very high R-squared values, large and significant t-values, and p-values less than 0.0001. Given that all of the regressions have p-values less than 0.0001, the change in the t-values for the coefficient on return forecast score (rank-order number) and the change in R-squared are the best indicators of the effect on the cross-section of imposing additional control constraints, especially in terms of the extent to which we are obtaining a return forecast dependency that is better isolated from the effect of non-forecast variables. Thus, we focus most of our attention here on the changes in the t-values and R-squared values as we impose different sets of control constraints.

The most significant structural feature is the jump in R-squared values and t-values when we impose the tax controls alone or impose the tax controls along with the three systematic risk factors. In particular, we observe that for the regressions in which forecast score is the explanatory variable, the imposition of the three risk controls alone produces a t-value on the slope coefficient of 6.44. Imposing the three tax controls alone produces a t-value of 17.81. The control set with both systematic risk and tax controls has a t-value of 18.97.

10.9.2 The Cross-Section of Realized Standard Deviations

Exhibit 10.8 summarizes linear regressions of the realized cross time standard deviations on both return forecast score and portfolio relative rank-order number. As with the regressions summarized in Exhibit 10.7, the results using return forecast score as the dependent variable are very similar to the results when we use portfolio rank-order number as the dependent variable.[25]

The highly significant intercept value simply reflects the ability to explain the positive average value for standard deviation. The coefficient on the linear term C1 is insignificant until tax controls (FL, EP, and DP) are imposed. The large jump in the t-value with the imposition of tax controls alone or in combination with other variables again indicates that controlling for tax effects is critical to isolate return

[25] In the LP version of the mathematical assignment program, we used rank-order differences as approximations for return score differences from average portfolio return scores. The fact that the regressions on longrun average returns and realized standard deviations are essentially the same in Exhibits 7 to 9 when we use rank-order number and return forecast score as alternative explanatory variables is evidence that this approximation is reasonable for the control-matched portfolios produced by the mathematical assignment program.

forecast performance from other distorting return factors. For the cross-section of realized standard deviations, it appears systematic tax effects are the most pertinent set of control variables rather than the usual systematic risk variables.

The plots of the long-run realized standard deviation in Exhibit 10.5 indicated that the tails of the rank-ordering had greater standard deviation than the middle of the distribution. For this reason, we are adding a quadratic term in assessing how well the standard deviation is explained by return forecast score. These are summarized in Exhibit 10.9. For all the regressions with both a linear and a quadratic term summarized in Exhibit 10.9, the coefficient on the linear term is significantly negative and the coefficient on the quadratic is significantly positive. The p-values are all less than 0.0001. This result supports the impression from the plots in Exhibit 10.5, especially Exhibits 10.5F on, that the cross-section of realized long-run standard deviations does not have a monotonic increase in realized standard deviation risk but rather has only an increase in each tail of the cross section.

10.9.3 The Cross-Section of Realized Skewness Coefficients

Exhibit 10.10 is a cross-sectional regression on the long-run realized skewness as measured by the skewness coefficient. The skewness is somewhat significant even with no control constraints. Adding just the beta control alone increases the t-value to 6.43 from just 3.86 with no control constraints. This jump in the t-value from imposing the beta control alone suggests that controlling for market movements by means of the beta control greatly increases the isolation of nonsystematic skewness from any market skewness. Interestingly, neither the size control alone nor the book-to-market control alone significantly changes the t-value of the coefficient on the skewness measure. However, the three systematic risk controls together increase the t-value from 6.40 for just beta controls to 7.13.

In contrast to the regressions for realized standard deviation, tax controls do not significantly change the regression slopes or t-values. In effect tax controls alone do not seem to help isolate nonsystematic skewness. However, the three risk controls plus the three tax controls together do increase the coefficient and the t-value somewhat.

Adding our two growth controls and other company-specific controls does increase significantly the ability of return forecast score to explain the cross-sectional skewness coefficient. We conclude that isolating non-systematic value-related skewness associated with the illustrative value-focused return forecasting model requires that we isolate these value-related return and risk effects from growth in particular.

In the final set of controls in Exhibit 10.9, we add two more model controls namely the cash-price ratio and the sales-price ratio. Given that the control set already contains the book to market ratio and the earnings price ratio, the net effect of this final control set is to remove from the cross-section the contribution to realized returns of all four current value ratios.

This final set of controls in Exhibit 10.9 gives a forecast response for just the relative value ratios. It is interesting that the t-value is the greatest for this set of control constraints. This high t-value for the response of realized skewness to the relative value subset of the eight forecast variables is strong evidence that the significant cross-sectional dependency of skewness on return forecast score is primarily from the relative value ratios rather than the current value ratios.

Since the relative value ratios measure attractiveness for a company relative to its own past value ratios, this result suggests, or at least is consistent with, company values returning to a moving mean.

10.10 Resolving Alpha Performance from Greater Risk Bearing

Having assessed how the cross-section of realized long-run returns, realized standard deviations, and realized skewness coefficients depends on the long-run average value of the return forecast score (portfolio rank-order number), we turn now to a regression test to address the critical question of the extent to which the realized cross-sectional return dependency is an alpha performance or a standard deviation risk effect, or possible a third moment effect. We recognize of course that the cross-sectional realized return dependency can be a combination of alpha performance, of greater SD risk, and even possibly third or higher moment effect. The crux of the alpha-risk resolution is to assess the extent to which a well-controlled (presumably, well-isolated) cross-section of long-run realized returns is explained by the return forecast score or by realized standard deviation or by higher moment measures such as skewness in our tests.

Exhibit 10.12 is a stepwise regression for every set of control variables that assesses in a stepwise fashion how well realized returns are explained by the return forecast alone, then a combination of the return forecast and the realized standard deviation, and finally, a combination of the return forecast score, realized standard deviation, and realized skewness coefficient. Given our assessment of significant nonlinearities in the cross-section of both realized standard deviation and realized skewness, we add both a linear and a quadratic term for both standard deviation and skewness in the stepwise tests.

We have already established in Exhibits 10.7 and 10.8 that long-run realized returns are well explained by both the return forecast score and the portfolio rank-order number. Adding first standard deviation and then skewness to these cross-sectional regressions in Exhibit 10.12 assesses the extent to which realized standard deviations explain the long-run average realized returns in the realized return cross-section in competition with average return forecast score. Adding both realized standard deviation and realized skewness assesses the extent to which these two variables jointly explain at least some of the long-run realized return cross-section.

While we have included stepwise regressions for all the sets of control constraints, the control constraint sets that are most pertinent to assessing the extent to which a well-isolated cross-sectional return dependency is a return to greater realized SD risk rather than an alpha performance effect are those controls sets that include at least systematic risk and tax controls if not risk, tax, and growth controls. In effect, for assessing whether the long-run realized return response to the forecast rank-ordering is an alpha performance effect or just a return to greater risk bearing, we call readers attention especially to the more complete sets of control variables, especially controls sets F and higher.

For control set F and beyond (that is for all the control sets that include both systematic risk and tax controls), the coefficient on the linear standard deviation variable C_2 is significantly negative. Coefficient C_3 (the coefficient on the quadratic standard deviation variable) is significant and positive. The quadratic term is reflecting the nonlinear, non-monotonic dependency that we have already noted in looking at the plots of standard deviation vs. portfolio rank-order number. The significant quadratic dependency reflects the fact that there is more standard deviation in each end of the cross-section.

The final panel in Exhibit 10.12 adds explanatory variables for skewness, namely a linear and quadratic term, each measured relative to the mean value of skewness in the cross-section. While Exhibits 10.10 and 10.11 showed that there was significant skewness in the cross-section of realized returns, the final panel in Exhibit 10.12 indicates that the long-run realized skewness values have very little ability to explain the long-run realized values of return.

10.10.1 Stepwise Regressions: Conclusions

The following summarizes the results of the stepwise regressions tabulated in Exhibit 10.12.

1. The long-run average realized returns have a significant dependency upon forecast return score, especially for control sets F and beyond.
2. The linear coefficient on standard deviation is generally negative and insignificant when there is just a linear standard deviation term in the cross-sectional regression. When there is also a quadratic term in the regression, the coefficient on the linear term becomes negative and significant. Either an insignificant or negative coefficient on standard deviation suggests that the return realizations are not explained by increasing risk as measured by realized standard deviation.
3. The fact that the skewness coefficient has very little incremental explanatory value beyond the forecast return score suggests that the realization of greater right skewness with increasing return score does not explain the apparently significant potential for realizing alpha performance from utilization of the forecast.

Given that investors generally like and presumably price right skewness, it is puzzling that realization of significant positive skewness is seemingly not priced in the realized return response to the forecast score. One explanation is that ex post realizations were different from ex ante expectations in at least some periods over the very long time period in this skewness assessment. Another is simply the difficulty of detecting nonsystematic skewness without the kind of return impact isolation methodology used in this study.

Because standard deviation does not explain most of the performance potential inherent in the dependency of long-run realized returns on return forecast score and given that there is additional positive skewness, the evidence for a positive alpha performance on the return response subsurface is even stronger with the positive skewness than would be the case if there were just this magnitude of returns without any positive skewness in the high return rank portfolios. Overall, the regressions summarized in Exhibit 10.12 indicate that for the cross-sections for which the return forecast as well isolated from other return impact variables, the significant long-run realized return dependency on the forecast is primarily an alpha performance potential rather than a systematic risk effect.[26]

10.11 Sharpe Ratios for the Upper Quartile PFO the Upper Quintile Portfolio Compared with Market Indices

The regressions summarized in Exhibit 10.12 indicate the long realized return is primarily an alpha performance effect. However, the tabulation and plotting of realized standard deviation in Exhibit 10.5 indicate that at least for the top three if not the top six portfolios, there is greater realized standard deviation. Thus, for the top ranked portfolios, it is pertinent to test further the question whether the greater realized return more than compensates for greater standard deviation risk. To address this question, we look at Sharpe ratios for the upper fractile portfolios for the overall time period and four major subperiods.

[26] Given that we have not done a formal assessment of trading cost and price impacts, we have generally said *alpha performance potential* in drawing conclusions from the performance information in the tests reported here. An ability to go from our assessment of alpha performance potential to likely *performance value* is supported by complementary research reported in Guerard et al. (2009) return forecast model that is an extension of the eight variable Bloch et al. (1993) model tested here. They test a ten-variable model that uses the eight variables of Bloch et al. (1993) plus a Fama–French momentum variable plus a variable representing a measure constructed from consensus I/B/E/S forecast revisions, breadth, and earnings yield. They report a total managed return of 12.52% with a standard deviation of 20.49%. The total active return is 8.7% ($t = 2.78$). Asset selection is 7.49% and the t-value on asset selection is 3.28. The Guerard, Menchero, and Miller universe consisted of all I/B/E/S, Compustat and CRSP stocks for January 1998 to December 2007 targeting the Russell 3000 growth index and subtracting 125 basis points for transaction costs each way. Turnover was constrained at 8% per month. It cost the portfolio about 300 basis points of return annually. What do the Guerard and Chettiappan (2009) results tell us about going from performance potential to performance value? Even though their test period is shorter than the 1967–2004 period used here, their results indicate that using the widely used BARRA multifactor performance assessment framework for evaluating forecast models of this type involves reductions of 300 to 500 basis points and thus suggests longrun performance value for the forecast model of Bloch et al. (1993) tested here even if we were to impose adjustment costs as large as 500 to 600 basis points annually.

The obvious way to use the value forecasts is to input the return forecasts to a mean–variance optimizer designed to match the index values of each control variable. The result would be a portfolio in each time period that optimized predicted return relative to past security-level variance and co-variance including especially security-level covariance with the value effect. This would optimize the predicted return relative to predictions of controllable risk. Given a forecast, the optimizer should reflect statistical confidence in both the return and the variance–covariance forecast.

Rather than picking the optimal trade-off of predicted return for predicted uncertainty and forecast confidence, we propose looking at a much weaker use of value-based return forecasts. We construct here a naïve "*upper fractile*" portfolio formation strategy that represents a lower bound on the achievable Sharpe ratio from value-based portfolio construction. Using a lower-bound (i.e., worst-case) value-based portfolio strategy to test the null hypothesis of no value means that we minimize the possibility of falsely rejecting the null hypothesis. Thus, using the weakest test of the comparative performance benefits of the value forecast means that we minimize the possibility of falsely rejecting a correct null hypothesis (Type 1 error). For this forecast methodology, the set of controls, and tests on 30 matched portfolios, we maximize statistical power for avoiding false rejection of a correct null hypothesis by using the weakest case of value-based performance.

We define the *upper quintile portfolio* as the top six portfolio fractiles (i.e., portfolio #24 to portfolio #30). We form it simply by combining the top 6 portfolios. The key point is that there is no use of an optimization algorithm beyond that used to construct the control-matched cross-section. Given that each of the top six portfolios in the cross section matches the ex ante population average values for each of the controls, the combination will continue to match the population averages for each controls.

The upper quintile uses only value-based forecast information and control values. As a simple portfolio formation rule that uses only return forecast and control variable information, finding significant performance benefits for the upper quintile portfolio is a weak test of the null hypothesis. If return forecast and control value information alone provides statistically and economically significant performance value, then we know any rule adding past risk and forecast uncertainty would do better. Thus, our construction of portfolios using only security-level value-based return forecasts and security-level ex ante control values would do even better. In effect, for the given portfolio size of approximately one fifth, the upper quintile portfolio is the least information (lowest Sharpe ratio) portfolio.

Exhibits 10.13 to 10.15 summarize Sharpe ratios for the upper quintile portfolio, for the lower quintile portfolio, and for both the CRSP value-weighted and equal-weighted indices. In addition to the overall 1967–2004 time period, we also provide Sharpe ratio comparisons for four subperiods:

- 1967–1974 (8 years)
- 1975–1985 (10 years)
- 1985–1994 (10 years)
- 1995–2004 (10 years).

For each of these time periods we have produced the Sharpe ratios for three sets of control constraints:

- No Controls (Exhibit 10.13)
- Beta, Size, FL, DY, Sag5, Sug3 (Exhibit 10.14)
- All Controls except ROI (Exhibit 10.15).

Focusing on the upper quintile portfolio vs. the CRSP equally weighted index for the case of all controls (Exhibit 10.15), the Sharpe ratios are 0.204 and 0.138 respectively for the overall time period. For the case of all controls in Exhibit 10.15, we note that the Sharpe ratios are larger than either of the CRSP indices in every sub-period.

Next we note that the upper quintile portfolio is consistently superior to the lower quintile portfolio. These indicate significant value to the return forecast score for producing portfolios that have superior long-run returns relative to realized long-run risk as measured by cross-time standard deviation. The favorable cross-sectional increase in skewness adds to the assertion of forecast value.

10.12 Synthesis, Conclusions, and Further Research

This research used an objectively specified return forecasting procedure adapted from Bloch et al. (1993). It is an adaptive updating of the relative weighting of eight fundamental ratio variables. The use of both current ratios and current ratios relative to the average of past values tends to reflect the two judgmental uses of value-ratios by value-focused analysts according to practiced-focused works such as Stowe et al. (2002). The objectively measured forecast score of each stock was used to rank order all non-financial stocks listed for at least 5 years with data in the intersection of the CRSP-COMPUSTAT databases.

By applying the forecast model each month from the start of 1967 to the end of 2004, we generated a time series of 30 forecast-ranked fractile portfolios. By applying control constraints in a step-wise fashion, the initial cross-sections were progressively transformed into related control-matched cross sections. By imposing controls in stepwise fashion with progressively more complete sets of controls, we were able to assess how imposing each set of controls changed the apparent portfolio-level dependency of both realized returns and realized standard deviations on the return forecast.

The cross-section of realized returns rank-ordered well with the return forecast score for the 38-year average and for each of four subperiod averages (as well as for individual years, which is not reported here because of space limitations). The long-run average return spread (high minus low) generally exceeds 8% annualized for the uncontrolled starting cross-section. Adding controls (isolating the forecast from the controls) actually increases the range of predicted returns. Given that these variables have such large magnitudes after removing any correlation with either the book-to-market variable or earnings yield and after controlling for beta, size, taxes, financial structure, and growth, and given that the majority of this cross-sectional return is not explained by the cross-sectional variation in ex post standard deviation, the value-based cross-sectional return dependencies are logically viewed as return anomalies in the sense of a significant cross-sectional return dependency that is not explained by risk.

Even when very high transaction costs are assessed for annual portfolio turnover for the upper quintile portfolio, the risk-adjusted, tax-neutralized, growth neutralized performance value relative to both the CRSP value-weighted and equal-weighted indices is economically large. For instance, an average annualized average return difference of more than 5% with comparable ex post standard deviations. This dominance is reflected in comparative Sharpe ratios.

In terms of cross-sectional return assessment methodology, using optimization to generate a cross-section of control-matched portfolios provides measurement error mitigation while controlling almost completely for the usual confounding of results that arises from magnification of low-level sample correlation in grouping observations. In terms of assessing a conditional functional dependency, the imposition of matched controls tends to ensure that the portfolio observations belong to a well-defined response subsurface.

Although our primary focus in this paper has been to illustrate variable isolation and return-risk assessment methods, our results in testing Bloch et al. (1993) would strongly reject a null hypothesis of no risk-adjusted, tax-corrected performance potential for this forecast model for the 1967–2004 time frame. While caution is always in order in drawing performance conclusions before testing transaction costs, price impacts, and any backtest or survival bias, the magnitudes and significance of the well-isolated return-forecast dependency seem to reject strongly the null of no performance potential even after allowing for all the necessary corrections.

Interestingly, the turning-point focused value forecast techniques modeled in this paper are significantly negatively correlated with return momentum. If significant risk-corrected, tax-corrected performance is a puzzle, we may have another performance puzzle to go with return momentum.

A second puzzle meriting further research is to explain why a model that simulates a subset of the assessment techniques of value-focused portfolio managers seems to do much better than the actual performance of most value managers. Transaction costs including trade impact on price is one possible explanation. A possible behavioral explanation is the difficulty in using well all eight fundamental value variables and/or the difficulty of judgmentally identifying alpha value at the stock level given the high stock-level noise-to-signal ratio in the forecasts.

The preceding summary-synthesis has focused primarily on what we have learned about forecast potential for the test forecast model of Bloch et al. (1993). Our primary purposes were to illustrate forecast isolation and alpha-risk resolution methods. In terms of methodology contributions we believe that this paper has accomplished the following:

- Established value for the response surface–subsurface optimization methods used to transform a rank-order-based cross-section of fractile portfolios into an associated control-matched cross section with the impact of controls removed from the cross section.

- Established that stepwise imposition of control constraints is a useful procedure for assessing return-forecast distortion and ultimately indicating whether a cross-sectional return dependency is well isolated from other return impact variables.
- Established that a time series of control-matched return cross-sections can be evaluated for alpha performance vs. risk compensation by using realized standard deviation (or other pertinent measures of realized risk such as semi-standard deviation) to assess/test the relative ability of the return forecast and the realized risk measure(s) to explain the cross-section of long-run realized returns.
- Expanded the set of firm-specific variables required to explain the cross-section of returns including especially the need to reflect systematic tax effects.
- Expanded our ability to explain/control for the effect of growth on the cross-section of returns by adding *sustainable growth* to the set of pertinent growth variables/controls.

In formulating a general return dependency on J-firm specific variables, we argued that the number of firm-specific variables required to isolate well the illustrative forecast model of Bloch et al. (1993) was large. The significant change in the return, standard deviation, skewness, and Sharpe ratio cross-sections with additions and variations in controls shows that the number is in fact at least as large as the number of control variables used here.

Among opportunities for further research are a need to study further the apparently large tax effect in the cross sections researched here, to improve further growth measures/controls, to relate the controls used here to other return impact variables, to study further the relation of the value of variables in Bloch et al. (1993) to return momentum, to research further the ability to use fundamental firm attributes to predict skewness including of course an explicit treatment of systematic coskewness, and especially to pursue further the question of how many variables are required to characterize completely the cross-section of returns including an expansion from the firm-specific explanatory variables used in this study to include as well both macro and monetary variables. Especially pertinent is the need to assess better how the significant performance potential identified in this research for the return forecast model of Bloch et al. (1993) [and presumably similar fundamental value forecast models] can translate into actual performance value by reflecting both trading costs and trade impact on market prices as well as doing the data mining corrections required for any backtest of the type reported here.

Acknowledgments The development of the optimization algorithm used to create the cross-sections of matched portfolios was supported by IBM (International Business Machines, Inc.) and the Utah Super Computer Institute. Both the Harold F. Silver Fund and Madeline Ruth Silver Fund at BYU provided data base support (CRSP and Compustat) as well as extensive research support for Bernell K. Stone. Stone also acknowledges Intel for their gift to Brigham Young University of two research computer servers. Greg Adams, and Troy Carpenter of the BYU finance research support center have provided exceptional data and statistical support. The authors acknowledge Mustafa Gultekin for past joint work, extensive interaction on the early stages of this work, and many helpful comments. Earlier versions of this research have been presented at Stern School Conference in Honor of Martin Gruber's 65th Birthday (New York University, November 2004), Environmental Protection Agency Workshop on Capital Markets (Laguna Beach, CA, February, 2006), 27th International Forecasting Symposium (New York, June 2007), INFORMS Conference on Optimization

(Atlanta, March 2008), and the 28th International Symposium on Forecasting (Nice, France, June 2008). Of course, the authors are responsible for any remaining errors or deficiencies.

References

Adrian, T. and J. Rosenberg. 2008. "Stock Returns and Volatility: Pricing the Short-Run and Longrun Components of Market Risk." *Journal of Finance* 63, 2997–3030.

Banz, R. 1981. "The Relationship Between Returns and Market Value of Common Stocks." *Journal of Financial Economics* 9, 3–18.

Basu, S. 1974. *The Relationship Between Investment Performance of Equity Securities and Their Price-Earnings Ratios: A Test of the Efficient Market-Expected Return Model*. Unpublished Doctoral Dissertation, Cornell University.

Basu, S. 1977. "Investment Performance of Common Stocks in Relations to Their Price Earnings Ratios: A Test of Market Efficiency." *Journal of Finance* 32, 663–682.

Basu, S. 1983. "The Relationship Between Earnings Yield, Market Value and Return for NYSE Common Stocks: Further Evidence." *Journal of Financial Economics* 12, 129–156.

Bernhard, A. 1959. The Evaluation of Common Stocks. New York: Simon and Schuster.

Bloch, M., J.B.Guerard Jr., H.M. Markowitz, P. Todd, and G.L. Xu. 1993. "A Comparison of Some Aspects of the U.S. and Japanese Equity Markets." *Japan & the World Economy* 5, 3–26.

Blume, M. 1980. "Stock Return and Dividend Yield: Some More Evidence," *Review of Economics and Statistics* 62, 567–577.

Blume, M.E., M.N. Gultekin, and N.B. Gultekin. 1990. "Validating Return-Generating Models." Working Paper.

Box, G.E.P. 1954. "The Exploration and Exploitation of Response Surfaces." *Biometrica* 10, 16–60.

Box, G.E.P. and N.R. Draper. 1959. "A Basis for the Selection of a Response Surface Design." *Journal of the American Statistical Association* 54, 622–654.

Box, G.E.P. and N.R. Draper. 1987. *Empirical Model Building and Response Surfaces*. New York: Wiley Interscience.

Box, G.E.P. and N.R. Draper. 2007. *Response Surfaces, Mixtures, and Ridge Analysis*. New York: Wiley Interscience.

Box, G.E.P., J.S. Hunter, and W.G. Hunter. 2005. *Statistics for Experimenters*. New York: Wiley Interscience.

Brennen, M. 1970. "Taxes, Market Valuation and Financial Policy." *National Tax Journal* 23, 417–429.

Brown, S.J. 1993. "Nonlinear Systems Estimation: Asset Pricing Model Applications." In H. Varian (ed.), Economic *and Financial Modeling with Mathematica*. New York: Springer Verlag.

Brown, S.J. 2008. "Elusive Return Predictability: Discussion." *International Journal of Forecasting* 24, 19–21.

Brown, S.J. and M.I. Weinstein. 1983. "A New Approach to Testing Asset Pricing Models: The Bilinear Paradigm." *Journal of Finance* 38, 711–743.

Campbell, J.Y., J. Hilscher, and H. Szilagyi. 2008. "In Search of Distress Risk." *Journal of Finance* 63, 2899–2939.

Chan, L.K.C., N. Jegadeesh, and J. Lakonishok. 1996. "Momentum Strategies." *Journal of Finance* 51, 1681–1713.

Conner, G., and R.A. Korajczak. 1988. "Risk and Return in an Equilibrium APT: Application of a New Test Methodology." *Journal of Financial Economics* 21, 255–289.

Conner, G. and R.A. Korajczak. 1993. "A Test for the Number of Factors in an Approximate Factor Model." *Journal of Finance* 48, 1263–1291.

Cooper, M.J., H. Gulen, and M.J. Schill. "Asset Growth and the Cross-Section of Stock Returns." *Journal of Finance* 63, 1609–1651.

Dhrymes, P.J., I. Friend, and N.B. Gülteikin. 1984. "A Critical Reexamination of the Empirical Evidence on the Arbitrage Pricing Theory." *Journal of Finance* 39, 323–346.

Dhrymes, P.J., I. Friend, M.N. Gültekin, and N.B. Gültekin. 1985a. "New Tests of the APT and Their Implications." *Journal of Finance* 40, 659–674.

Dhrymes, P.J., I. Friend, N.B. Gültekin, and M.N. Gültekin. 1985b. "An Empirical Examination of the Implications of Arbitrage Pricing Theory." *Journal of Banking and Finance* 9, 73–99.

Dremen, D. 1979. *Contrarian Investment Strategy*. New York: Random House.

Dremen, D. 1998. *Contrarian Investment Strategies: The Next Generation*. New York: Simon and Schuster.

Elton, E.J., M.J. Gruber, and M.W. Padberg. 1979. "Simple Criteria for Optimal Portfolio Selection: The Multi-Index Case." In E.J. Elton and M.J. Gruber (eds.), *Portfolio Theory, 25 Years After: Essays in Honor of Harry Markowitz*. Amsterdam: North-Holland.

Fama, E.F. and K.R. French. 1988. "Dividend Yields and Expected Stock Returns." *Journal of Financial Economics* 22, 3–25.

Fama, E.F. and K.R. French. 1992. "Cross-Sectional Variation in Expected Stock Returns." *Journal of Finance* 47, 427–465.

Fama, E.F. and K.R. French. 1995. "Size and the Book-to-Market Factors in Earnings and Returns. *Journal of Finance* 50, 131–155.

Fama, E.F. and K.R. French. 2008a. "Dissecting Anomalies." *Journal of Finance* 63, 1653–1678.

Fama, E.F. and K.R. French. 2008b "Average Returns, B/M, and Share Issues," *Journal of Finance* 63, 2971–2995.

Fama, E.F. and J. MacBeth. 1973. "Risk, Return, and Equilibrium: Empirical Tests." *Journal of Political Economy* 81, 607–636.

Ferson, W.E. and C.R. Harvey. 1991a. "The Variation of Economic Risk Premiums." *Journal of Political Economy* 99, 385–415.

Ferson, W.E. and C.R. Harvey. 1991b. "Sources of Predictability in Portfolio Returns." *Financial Analysts Journal* 47, 49–56.

Ferson, W.E. and C.R. Harvey. 1995. "Explaining the Predictability of Asset Returns." In A Chen (ed.), *Research in Finance* 11.

Ferson, W.E. and C.R. Harvey. 1998. "Fundamental Determinants of National Equity Market Returns: A Perspective on Conditional Asset Pricing." *Journal of Banking & Finance* 21, 1625–1665.

Graham, B. and D.L. Dodd. 1934. *Security Analysis*. New York: McGraw-Hill Book Company.

Graham, B., D. Dodd, and S. Cottle. 1962. *Security Analysis: Principles and Techniques*, 4th edition. New York: McGraw-Hill Book Company.

Guerard, J.B. Jr. and S. Chettiappan. 2009. "Stock Selection Modeling and Data Mining Corrections: Long-Only versus 130/30 Models." This Volume.

Guerard, J.B. Jr., M.G. Gultekin, and B.K. Stone. 1997. "The Role of Fundamental Data and Analysts' Earnings Breadth, Forecasts, and Revisions in the Creation of Efficient Portfolios." In A. Chen (ed.), *Research in Finance* 15, 69–91.

Hamada, R.S. 1972. "The Effect of the Firm's Capital Structure on the Systematic Risk of Common Stocks. *Journal of Finance* 27, 435–452.

Haugen R.A. 2001. *Modern Investment Theory*, 5th edition. Upper Saddle River, NJ: Prentice Hall.

Haugen, R.A. 1999. *The Inefficient Stock Market*, 5th Edition. Upper Saddle River, NJ: Prentice Hall.

Haugen, R.A. and N. Baker. 1996. "Communality in the Determinants of Expected Results." *Journal of Financial Economics* 41, 401–440.

Jaffe, J., D.B. Keim, and R. Westerfield. 1989. "Earnings, Yields, Market Values, and Stock Returns." *Journal of Finance* 44, 135–148.

Khuri, A.I., and J.A. Cornell. 1996. *Response Surfaces: Designs and Analysis*. New York: Marcel Dekker, Inc.

Lakonishok, J., A. Shleifer, and R.W. Vishny. 1994. "Contrarian Investment, Extrapolation, and Risk." *Journal of Finance* 49, 1541–1578.

Latane, H.A., D.L. Tuttle, and C.P. Jones. 1975. *Security Analysis and Portfolio Management*, 2nd Edition. New York: The Roland Press.

Lease, R.C., K. John, A. Kalay, U. Loewenstein, and O.H. Sarig. 2000. *Dividend Policy: Its Impact on Firm Value*. Boston: Harvard Business School Press.

Levy, H. 1999. Introduction to Investments. Cincinnati: South-Western College Publishing. Second Edition.

Markowitz, H.M. 1952. "Portfolio Selection." *Journal of Finance* 7, 77–91.

Markowitz, H.M. 1959. *Portfolio Selection: Efficient Diversification of Investment*. Cowles Foundation Monograph No. 16. New York: Wiley.

Markowitz, H.M. 1976. "Investment in the Long Run: New Evidence for an Old Rule." *Journal of Finance* 31, 1273–1286.

Markowitz, H.M. 1984. "The Two-Beta Trap." *Journal of Portfolio Management* 11, 12–19.

Markowitz, H.M. 1987. *Mean–Variance Analysis in Portfolio Choice and Capital Markets*. Oxford: Basil Blackwell.

Markowitz, H.M. 1991. "Foundations of Portfolio Theory." *Journal of Finance* 46(2), 469–477.

Markowitz, H.M. 2000. *Mean–Variance Analysis in Portfolio Choice and Capital Markets*. New Hope, PA: Frank J. Fabozzi Associates.

Markowitz, H.M., and G. Xu. 1994. "Data Mining Corrections." *Journal of Portfolio Management* 21, 60–69.

Myers, R.H. 1976. *Response Surface Methodology*. Boston: Allyn & Bacon.

Myers, R.H., A.I. Khuri, W.H. Corter Jr. 1989. "Response Surface Methodology: 1966–1988, *Technometrics*, 31, 137–157.

Peterson, P., D. Peterson, and J. Ang. 1985. "Direct Evidence on the Marginal Rate of Taxation on Dividend Income." *Journal of Financial Economics* 14, 267–282.

Pilotte, E.A. 2003. "Capital Gains, Dividend Yields, and Expected Inflation." *Journal of Finance* 58, 447–466.

Ramnath, S., S. Rock, and P. Shane. 2008. "The Financial Analyst Forecasting Literature: A Taxonomy with Suggestions for Further Research." *International Journal of Forecasting* 24, 34–75.

Rosenberg, B. 1974. "Extra-Market Components of Covariance in Security Returns." *Journal of Financial and Quantitative Analysis* 9, 263–274.

Rosenberg, B., and V. Marathe. 1979. "Tests of Capital Asset Pricing Hypotheses." In H. Levy (ed.), *Research in Finance* 1.

Rudd, A., and B. Rosenberg. 1979. "Realistic Portfolio Optimization." In E. Elton and M. Gruber (eds.), *Portfolio Theory, 25 Years After*. Amsterdam: North-Holland Publishing Co.

Rudd, A., and B. Rosenberg. 1980. "The 'Market Model' in Investment Management." *Journal of Finance* 35, 597–607.

Rosenberg, B., and A. Rudd. 1977. "The Yield/Beta/Residual Risk Tradeoff." Working paper 66. Research Program in Finance, Institute of Business and Economic Research, University of California, Berkeley.

Rosenberg, B., and A. Rudd. 1982. "Factor-Related and Specific Returns of Common Stocks: Serial Correlation and Market Inefficiency." *Journal of Finance* 37, 543–554.

Sharpe, W.F. 1963. "A Simplified Model for Portfolio Analysis." *Management Science* 9, 277–293.

Sharpe, W.F. 1964. "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk" *Journal of Finance* 19, 425–442.

Sharpe, W.F. 1967. "A Linear Programming Algorithm for Mutual Fund Portfolio Selection." *Management Science* 13, 449–510.

Sharpe, W.F. 1971. "A Linear Programming Approximation for the General Portfolio Problem." *Journal of Financial and Quantitative Analysis* 6, 1263–1275.

Stone, B.K. 1973. "A Linear Programming Formulation of the General Portfolio Selection Problem." *Journal of Financial and Quantitative Analysis* 8, 621–636.

Stone, B.K. 2003. "Rank-Based Security Grouping, Multicolinearity, and the Assessment of Cross-Sectional Return Dependencies," Working Paper. The Marriott School of Management, Brigham Young University.

Stone, B.K., D.L. Adolphson, and T.W. Miller. 1993. “Optimal Data Selection and Grouping: An Extension of Traditional Response Surface Methodology to Observational Studies Involving Noncontrolled Empirical Data Generation.” *Advances in Mathematical Programming and Financial Planning*. 3, 39–68.

Stowe, J.D., T.R. Robinson, J.E. Pinto, D.W. McLeavey. 2002. *Analysis of Equity Investments: Valuation*. Charlottesville, VA: Association for Investment Management and Research.

Stowe, J.D., T.R. Robinson, J.E. Pinto, D.W. McLeavey. 2007. *Equity Asset Valuation*. Hoboken NJ: John Wiley & Sons for CFA Institute and Association for Investment Management and Research.

Timmerman, A. 2008a. “Elusive Return Predictability.” *International Journal of Forecasting* 24, 1–18.

Timmerman, A. 2008b. “Reply to Discussion of Elusive Return Predictability.” *International Journal of Forecasting* 24, 29–30.

Williams, J.B. 1938. *The Theory of Investment Value*. Cambridge: HarvardUniversity Press.

Ziemba, W.T. 1991 “Fundamental Factors in US and Japanese Stock Returns.” Berkeley Program in Finance.

Chapter 11
Robust Portfolio Construction

R. Douglas Martin, Andrew Clark, and Christopher G. Green

11.1 Introduction

Classical mean–variance optimization (MVO) uses as its basic inputs an estimate of the mean returns and an estimate of the covariance matrix of the returns. It is well known that historical sample mean estimates have large relative variability and as such they are an Achilles heel of MVO, often resulting in poor diversification and instability of portfolio weights with respect to small absolute changes in the mean returns estimates. For this reason, portfolio managers typically use a presumably better alternative "alpha" forecast of mean returns, constructed in a proprietary manner or purchased from a commercial provider. Nonetheless we shall, for the sake of simplicity in introducing the use of robust methods for portfolio construction, assume that the sample mean of historical returns is used, keeping in mind that robust statistical methods may well be of use in building alternative alpha forecast models.

As for estimating variances and covariances, conventional wisdom is that the classical historical estimates have greater relative precision than classical sample mean estimates in situations where the number of returns is substantially larger than the number of assets. See for example Chopra and Ziemba (1993). Correspondingly, portfolio managers often assume that the estimation error introduced by these estimates is relatively negligible. The examples in this paper show that in the presence of outliers this need not be the case.

We note that a factor model is typically used for optimization and risk management of portfolios containing a large number of assets, one purpose of doing so being dimensionality reduction.[1] We finesse the need to treat such models in

R.D. Martin (✉) and C.G. Green
Department of Statistics, University of Washington, Seattle, WA, USA
e-mail: doug@stat.washington.edu

A. Clark
Thomson Reuters Indices, New York, NY, USA

[1] Assuming a fundamental factor model for an equity portfolio, other reasons are that (a) the risk factors chosen, e.g., book-to-market, size, leverage, momentum, industry sector, allow for decomposition of risk in terms of risk factors that investors are familiar with and (b) the risk factor model supports optimization with constraints on bets on such factors.

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_11,
© Springer Science+Business Media, LLC 2010

detail by focusing only on small portfolios, portfolios with a small number of assets, and discuss them briefly with respect to the difficulties posed by outliers in the risk factors.

It is a fact that asset returns have nonnormal fat-tailed skewed distributions to various degrees depending on the asset class and frequency, i.e., small-cap returns are more fat tailed than large-cap returns, hedge fund returns are more fat tailed and skewed than equity returns, and daily returns are more nonnormal than weekly returns, which are in turn more nonnormal than monthly returns. This awareness has led to considerable research literature on how to deal with such distributions in a portfolio performance enhancing manner. Two branches of research focus on the use of skewness and kurtosis in asset pricing models and portfolio optimization, respectively, and another branch of research focuses on the modeling and use of fat-tailed skewed probability distribution models.

One relatively recent skewness and kurtosis approach to portfolio optimization uses a Cornish–Fisher expansion (1960) to derive a "modified" value-at-risk (MVaR) risk measure for use in a mean vs. MVaR optimization variant of MVO. See for example Favre and Galeano (2002). Another approach is based on three- and four-moment CAPM-type factor models or generalized CAPM risk measures. See for example Kraus and Litzenberger (1976), Harvey et al. (2004), and Jurczenko and Maillet (2006).

A particularly attractive approach to portfolio optimization is to fit fat-tailed skewed distribution models, in particular stable distributions or skewed t-distributions, and combine such models with an expected tail loss (ETL) risk measure, a.k.a., conditional-value-at-risk (CVaR). This results in a mean-ETL approach optimization that is a highly attractive alternative to MVO. See for example Martin et al. (2003), Rachev et al. (2009) in this volume, and the introductory book by Rachev et al. (2005). For a definitive advanced treatment of Stable-Paretian distributions, see Rachev and Mittnik (2000).

It is a fact that outliers frequently occur in asset returns, where an outlier is defined loosely to be "an unusually large value well-separated from the bulk of the returns." While outliers are often well modeled by fat-tailed asset returns distributions, this is not always the case. Sometimes outliers occur in asset returns whose histories do not adequately support good fits of fat-tailed distributions, typically because of data sparseness or difficult-to-model regime shifts, or the data is multimodal and would best be served using mixture models. Also, outliers may sometimes be caused by data errors. Unfortunately outliers in a portfolio's asset returns often have a large adverse influence on the classical estimates of mean returns, covariances, and correlations, and hence on MVO efficient frontiers, as the reader will soon discover.

The main goal of this paper is to introduce and establish the value of using robust statistical methods that are not much influenced by outliers in a robust MVO method that is complementary to the classical MVO method. We emphasize that a primary use of the robust MVO approach is to diagnose likely adverse outlier influence on a classical MVO portfolio. Upon deeper examination of the nature and timing of the outliers, the robust MVO approach will sometimes be the preferred solution.

We stress that our use of the term “robust” means robust toward outliers, or equivalently, in terms of distribution models, it means delivering good performance near a normal distribution model where the nearby model can be fat tailed, skewed, and outlier generating. This is in contrast to an alternative use of the term “robust portfolio optimization” to describe the method of modifying the quadratic utility function with a quadratic penalty term for the uncertainty in mean returns estimates. This leads to an advanced optimization technique called second-order cone programming. See for example Ceria and Stubbs (2006) and related literature.

For another viewpoint on the Ceria and Stubbs approach see Scherer (2007), who has shown that for portfolio optimization with only a full-investment constraint the efficient frontier obtained by such an approach is equivalent to the efficient frontier obtained using a simple Bayes–Stein mean returns estimate approach. This begs the question of how much the relatively complex optimization cited above really buys the portfolio manager.

We note in passing that the word *robust* is a very resonant one that is sometimes used quite loosely, e.g., to mean an overall quality that holds up under varying conditions. See for example, Fabozzi et al. (2007), who use “Robust Portfolio Optimization” as a key part of the title while only a small fraction of the book deals with either of the two kinds of robust portfolio optimization cited above.

A search on “Robust Portfolio Optimization” reveals a large number of links to papers on the type of robustness described in the above paragraph and relatively a few links to papers on outlier robust portfolio optimization, among which are Lauprete et al. (2002), Gao (2004), Perret-Gentil and Victoria-Freser (2004), DiMiguel and Nogales (2006), and Welsch and Zhou (2007). See also Chernobai and Rachev (2006) for an application to operational risk.

The remainder of the paper is organized as follows. Section 11.2 provides a motivating example of returns outliers and their influence on classical sample mean, correlation, and covariance estimates for a four asset small-cap portfolio and introduces influence functions as a convenient tool for assessing outlier influence. Section 11.3 briefly discusses the basic concepts of robustness as developed in the statistics community. Section 11.4 introduces useful robust estimators of means, correlations, and covariances for application in constructing robust MVO portfolios. Section 11.5 compares classical and robust MVO portfolio efficient frontiers for portfolios of small-cap stocks, ETFs, commodities futures contracts, and small-cap stocks. Section 11.6 also shows how robust distances based on robust covariances, as well as visualization and clustering methods, may be used to reliably detect multidimensional outliers and in the exposures matrix of a fundamental factor model. Section 11.7 emphasizes the diagnostic value of robust portfolio construction methods. Section 11.8 discusses considerations in the use of robust mean–variance portfolio optimization vs. mean-ETL portfolio optimization based on fat-tailed skewed distribution models, and the complementary use of both methods. Section 11.9 discusses three other applications of robust statistics in portfolio management. Finally, Section 11.10 summarizes our results and states directions for future research.

11.2 Outliers in Asset Returns and Their Influence

In this section, we briefly describe four example portfolios of different asset classes that contain one-dimensional outliers in some of the individual asset returns as well as jointly in higher dimensions. Focusing on one of the portfolios, we illustrate the degree of influence that a small number of outliers can have on the classical mean, covariance, and correlation estimates used as the basic building blocks for MVO portfolio construction. Then we briefly discuss influence functions, a useful tool borrowed from the statistical literature for comparing alternative estimators based on their susceptibility to outlier influence.

11.2.1 Example Portfolios with Outliers

We consider small portfolios of four different asset classes in order to indicate the pervasiveness of outliers and their influence on classical estimates across distinct asset classes. The four portfolios are:

- Small-cap stocks (four stocks, monthly returns from 1/31/97 to 12/31/01)
- Hedge funds (eight hedge funds, monthly returns from 10/31/99 to 9/30/04)
- ETFs (eight ETFs, daily returns from 7/30/07 to 3/03/08)
- Commodities (eight contracts, monthly returns from 1/31/04 to 12/31/08)

The small caps were selected from the CRSP database, the ETFs were selected from the Lipper Optimal Indices, the commodities were selected from the Reuters-CRB index (a broad-based commodity index that has been in existence for more than 50 years), and the hedge fund returns were selected from a sample of hedge fund returns provided with the Cognity fund-of-funds portfolio construction and risk management product courtesy of FinAnalytica, Inc. (http://www.finanalytica.com). We chose small portfolios for the purpose of clear visual display of outliers, their influence on classical estimates, and subsequent results using robust methods. Subsequent study will be required to evaluate the use of the robust methods on large portfolios. In this section, we discuss the existence of outliers in the small-cap portfolio, supported by useful exploratory data analysis (EDA) visualization displays.

EDA displays for the hedge fund, ETF, and commodities portfolios are provided in the Appendix. Upon completion of reading this section, the reader is encouraged to browse the displays in the Appendix, as they reveal one- and two-dimensional outliers to various degrees in the asset returns for each of these three portfolios.

11.2.1.1 Small-Cap Returns

Figure 11.1 shows the time series and normal QQ-plots of the small-cap returns. The small-cap example is a relatively simple one with a few outliers that are easily detected in one-dimensional QQ-plot views of the data. However, it is important to know that in two dimensions and higher, outliers may reveal themselves more clearly in two-dimensional (and higher) views than they appear, if at all, in the one-dimensional time series and QQ-plot views. One can get a sense of this kind of behavior in the pairwise scatter plot for hedge fund returns in Fig. 11.2c. For example, in the scatter plot of F1 vs. F7, there are two clear two-dimensional

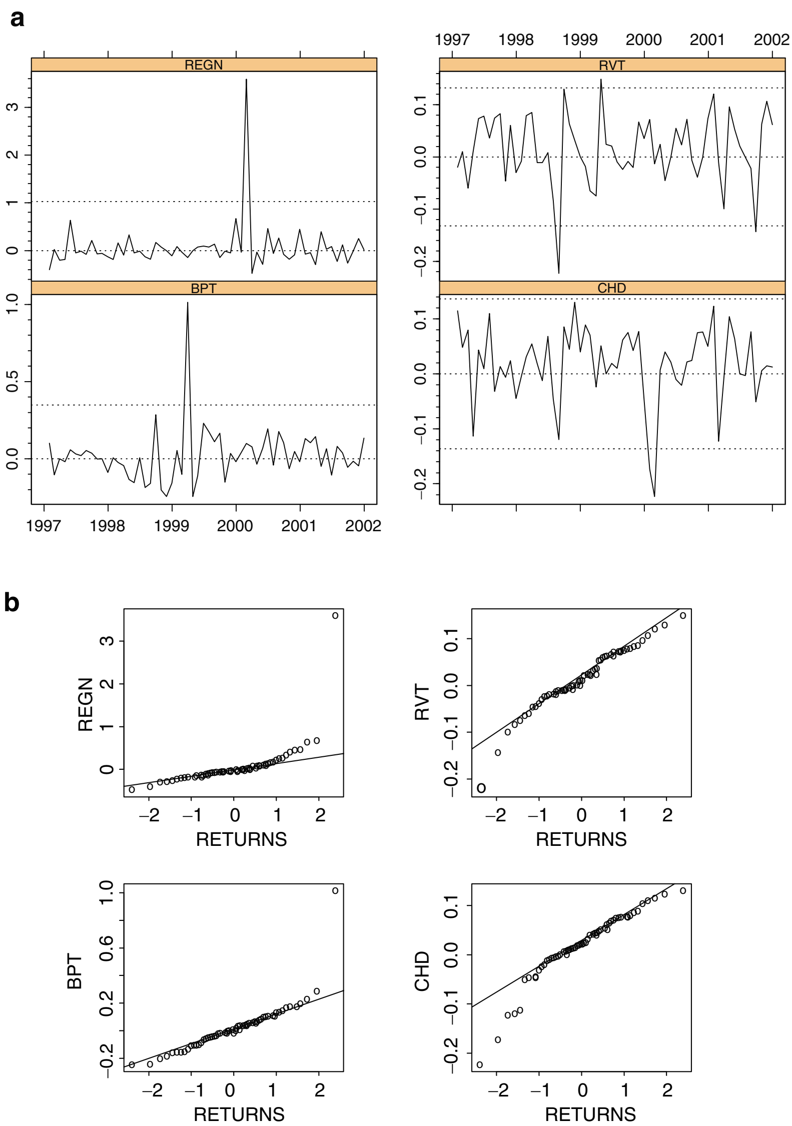


Fig. 11.1 (**a**) Time series of small-cap returns. (**b**) QQ-plots of small-cap returns. (**c**) Scatter plots of small-cap returns

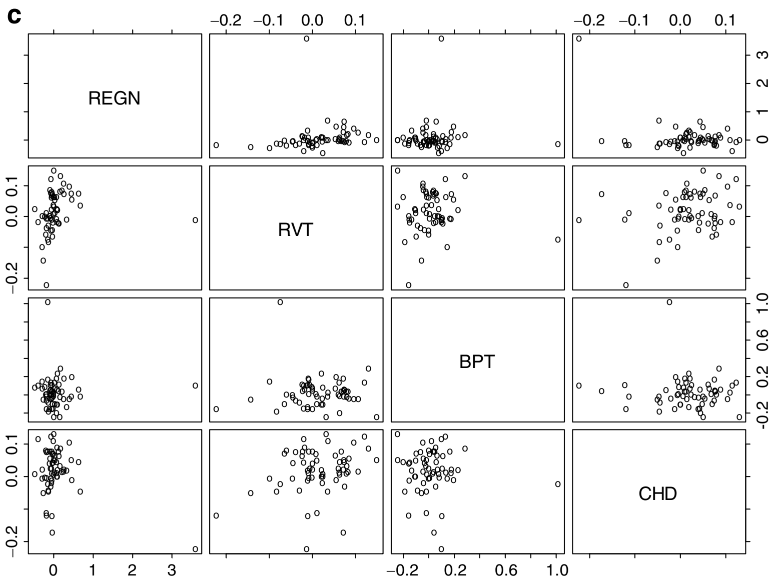


Fig. 11.1 (continued)

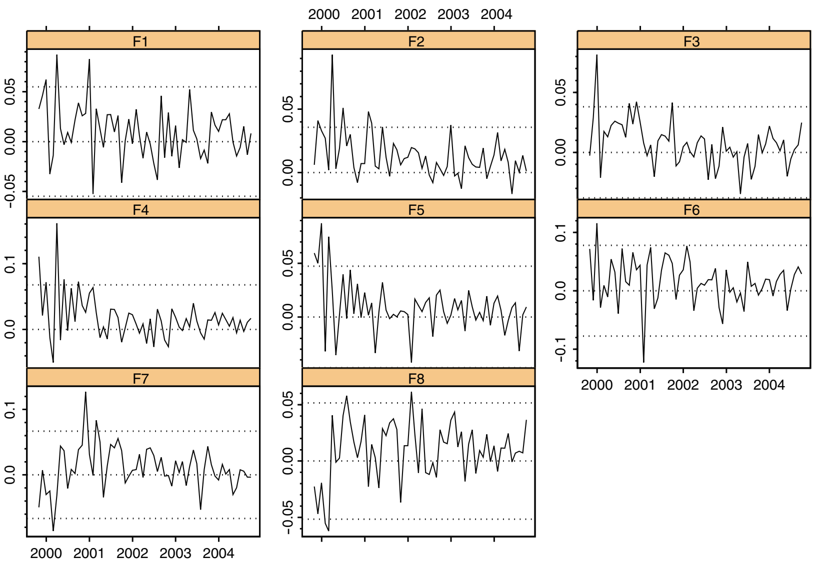
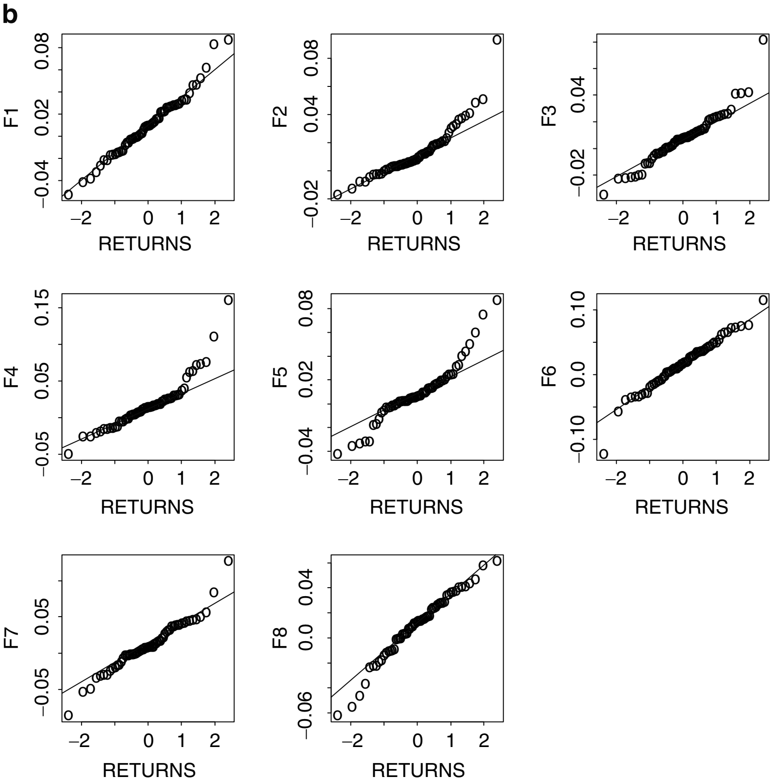


Fig. 11.2 (**a**) Time series of hedge funds returns. (**b**) QQ-plots of hedge funds returns. (**c**) Scatter plots of hedge funds returns. The time series clearly reveal chaotic behavior of some of the hedge fund returns in the first year, which encompasses the peak and collapse of the dotcom bubble, as well as a few outliers throughout the series. The QQ-plots both reveal fat-tailed and skewed non-normality to various degrees for all the hedge funds, with outliers clearly in evidence for F2, F3, F4, F6, and F7



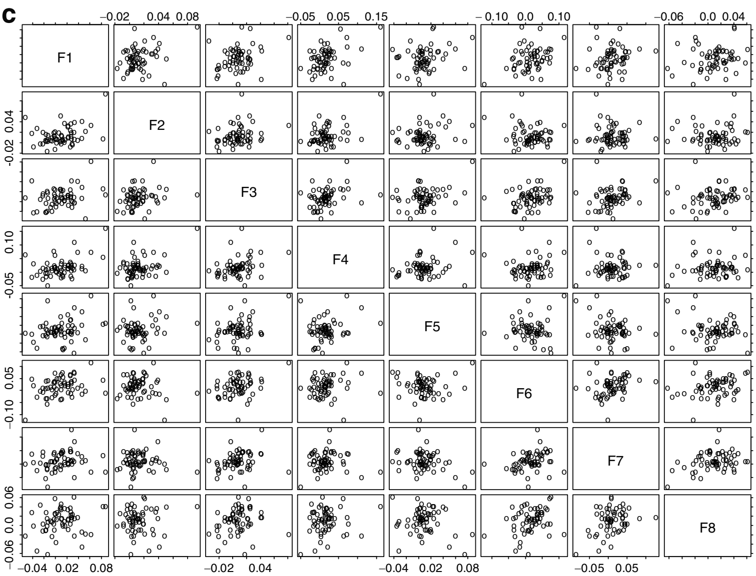


Fig. 11.2 (continued)

outliers in the upper left portion of the plot. In fact many two-dimensional outliers reveal themselves in many of the scatter plots more strongly than in the one-dimensional QQ-plot views. In general, none of the scatter plots are well modeled by an elliptical bivariate normal distribution. Note, however, that most of the scatter plots have a central bulk of the data that has an approximate elliptical shape. Similar comments to those above apply to corresponding plots for the ETF and commodities portfolios in the Appendix.

Three univariate normality tests were also conducted on the data. In Tables 11.1–11.4, we show the p values of these tests of normality – Jarque–Bera (JB), Shapiro–Wilk (SW), and Kolmogorov–Smirnov with the Lilliefors correction (KSL). The securities and time periods used are the same as those noted earlier and in Sect. 11.2.2. Significant differences from normality are displayed in bold font.

Table 11.1 Normality test results for small cap returns

Small-cap ticker	JB	SW	KSL
REGN	**0.0003**	**0.0000**	**<0.01**
RVT	**0.0210**	**0.0327**	<0.20
BPT	**0.0000**	**0.0000**	**<0.01**
CHD	**0.0003**	**0.0005**	**<0.05**

Table 11.2 Normality test results for hedge funds returns

Hedge fund ID	JB	SW	KSL
F1	0.3724	0.4401	> 0.20
F2	**0.0000**	**0.0000**	**<0.01**
F3	**0.0116**	**0.0101**	<0.15
F4	**0.0000**	**0.0000**	**<0.01**
F5	**0.0154**	**0.016**	**<0.05**
F6	0.1700	0.1691	>0.20
F7	0.0543	0.0557	<0.15
F8	0.1987	0.2050	>0.20

Table 11.3 Normality test results for ETF returns

ETF ID	JB	SW	KSL
INDIA	**0.0000**	**0.0001**	<0.20
Cons. discr.	**0.1250**	**0.0170**	**<0.05**
Sterling trust	**0.0000**	**0.0010**	**<0.05**
Yen trust	**0.0045**	**0.0038**	**<0.01**
Fin. sector	**0.0003**	**0.0001**	**<0.01**
Lehman 1–3 treas	**0.0103**	**0.0040**	**<0.05**
Lehman int. bond	**0.0000**	**0.0006**	**<0.01**
Energy sector	**0.0015**	**0.0008**	**<0.01**

Table 11.4 Normality test results for commodities returns

Commodities	JB	SW	KSL
Cattle	0.2020	0.1908	>0.20
Copper	**0.0145**	**0.0139**	>0.20
Coffee	0.1844	0.1834	<0.20
Hogs	0.1110	0.1110	<0.20
Silver	0.1476	0.1486	<0.20
Orange juice	0.5603	0.5582	>0.20
Platinum	**0.0000**	**0.0001**	**<0.01**
Crude oil	**0.0195**	**0.0200**	**<0.01**

11.2.2 Examples of Outliers Influence

Let us consider the influence of the outliers in Fig. 11.1 on the sample means, correlations, volatilities, and covariances.

11.2.2.1 Sample Mean Returns Estimates

The returns time series and QQ-plot of the stock with ticker REGN reveals one huge returns outlier whose value is 3.59. The monthly mean return estimate of REGN is 6.6%, unrealistically large from an investor point of view. On the other hand, the monthly mean return estimate with the outlier removed is 0.63%. From the perspective of a cleaned series with the outlier removed, the influence of the outlier is to increase the monthly mean return estimate by 6% (72% annually), which would be regarded as a gross distortion by any investor. Using a normal distribution fitted to the REGN returns without the outlier, the probability of observing a return at least as large as the outlier value 3.59 is zero to more than 20 decimal places.

The single outlier in the BPT returns series has the influence relative to the outlier-free returns of increasing the monthly mean returns estimate from 0.46 to 2.45%, an annual increase of about 24%, which would appear as grossly optimistic to a careful investor. The influence of the dominant negative outlier in the RVT returns has the effect relative to the outlier-free returns of decreasing the monthly mean returns estimate from 1.85 to 1.45%. The two most negative outliers in the CHD returns have the effect relative to the outlier-free returns of decreasing the monthly mean returns estimate from 2.52 to 1.78%.

Simple one-dimensional outlier treatment, such as use of trimmed or Winsorized means (see Sect. 11.3), could be used to avoid the influence of outliers on sample mean estimates of mean returns.

11.2.2.2 Sample Covariance and Correlation Estimates

The usual sample covariance matrix estimate is given by

$$\hat{\mathbf{C}} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{r}_t - \bar{\mathbf{r}})(\mathbf{r}_t - \bar{\mathbf{r}})',$$

where $\mathbf{r}_t = (r_{t1}, r_{t2}, \ldots, r_{tN})'$ is the column vector of returns at the t and $\bar{\mathbf{r}}$ is the sample mean of the $\mathbf{r}_t$, $t = 1, \ldots, T$ with asset sample means $\bar{r}_i$, $i = 1, \ldots, N$. $\hat{\mathbf{C}}$ has elements $\hat{C}_{ij}$, $i, j = 1, \ldots, N$ with diagonal elements $\hat{\sigma}_i^2$ the returns sample variance estimates. The estimate $\hat{\mathbf{C}}$ is a maximum-likelihood estimate in the case of normally distributed returns. It is slightly biased, and so the divisor T is often replaced by $T-1$ to obtain an unbiased estimate. The sample correlation coefficients $\hat{\rho}_{ij}$ are obtained from $\hat{\mathbf{C}}$

$$\hat{\rho}_{ij} = \frac{\hat{C}_{ij}}{\hat{\sigma}_i \hat{\sigma}_j}.$$

Unlike the case of the individual sample mean estimates in $\bar{\mathbf{r}}$, the extent to which the $\hat{C}_{ij}$ and the $\hat{\rho}_{ij}$ are influenced by outliers depends upon the multidimensional configurations of the outliers, not just their one-dimensional locations. The following brief discussion on subjective visual identification of two-dimensional outliers for the small-cap returns and their influence on the correlations is intended to give the reader a flavor for the nature of the multidimensional outlier returns problem.

In Fig. 11.1c, one would visually identify as outliers all those returns that are more or less clearly separated from a roughly elliptical central bulk of the returns. The single isolated, large positive one-dimensional outliers in the REGN and BPT returns continue to reveal themselves clearly as two-dimensional outliers in the scatter plots. The single negative one-dimensional outlier in RVT reveals itself in much the same way in the scatter plots of RVT vs. REGN and CHD returns scatter plots. However, in the scatter plot of RVT vs. BPT, the next two smallest RVT reveal themselves as potential mild outliers, i.e., clearly separated from the central bulk of the data, and this was not the case in the one-dimensional QQ-plot view of these returns. Finally, the scatter plots of REGN, RVT, and RBT vs. CHD returns reveal that it is not only the two most extreme negative returns that are outliers, but also the next three smallest returns as well, i.e., the five most negative returns in CHD are two-dimensional outliers.

Based on the above discussion we created a set of returns by modifying not only the single outlier returns in the REGN and BPT returns, but also the negative return in RVT and the five smallest returns in CHD as follows. From a one-dimensional point of view, this would involve modification of eight four-dimensional vectors of returns. However, there are two rows of the data matrix (with assets in columns and observations in rows) that contain two of the outliers each, so there are only six rows of returns to be modified. They were modified by replacing each row with the sample means 0.018, 0.021, 0.004, 0.034 of the asset returns with the rows

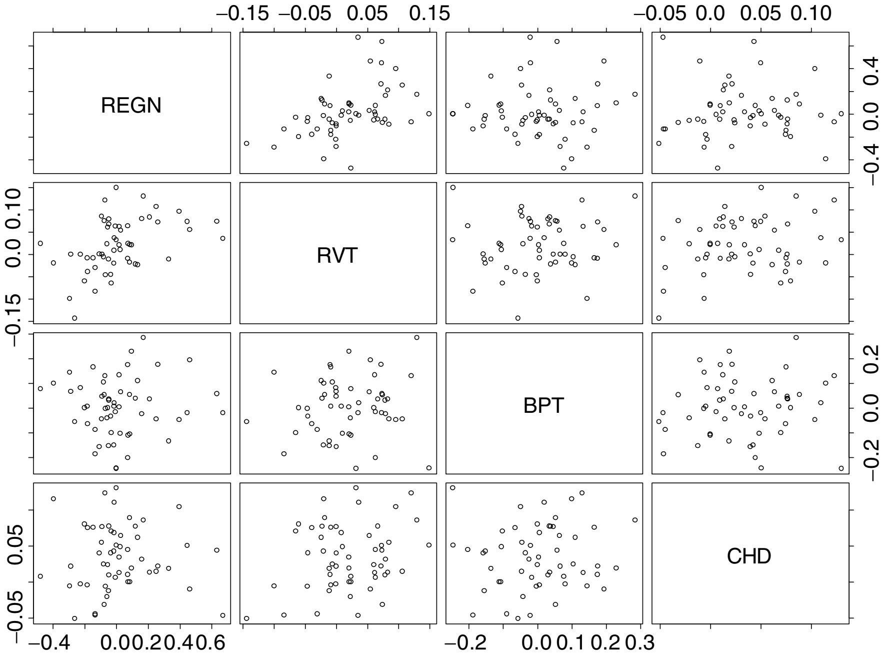


Fig. 11.3 Small-cap returns with outliers replaced by mean values

deleted. Figure 11.3 shows the scatter plots based on this modification and Fig. 11.4 shows the correlation values and ellipses computed from the original returns and the "cleaned" returns modified as above.

The differences in the correlations for the original and cleaned data are indeed substantial for the pairs (REGN, RVT) and (REGN, CHD). In the both cases, the outliers influence results in substantially smaller correlations than with the outliers removed, leading the portfolio manager to be overly optimistic about the diversification opportunities associated with these two pairs of small-cap stocks.

Computation of the volatilities of small-cap returns for the original and cleaned data and the ratios of the latter to the former result in:

	REGN	RVT	BPT	CHD
Original	0.51	0.07	0.17	0.07
Cleaned	0.21	0.06	0.11	0.04
Ratio	0.42	0.85	0.65	0.63

Thus elimination of outliers generally reduces the volatility (which we take to be standard deviation), i.e., it gives the portfolio manager an impression that the volatility risk in each stock is smaller than with the original data. However, one should keep in mind that volatility is a symmetric measure and therefore gives a misleading indication of risk when inflated by a large positive returns outlier.

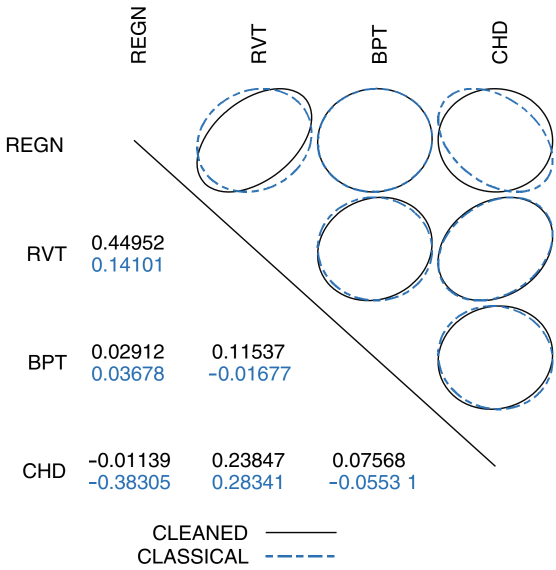


Fig. 11.4 Sample correlations for original (classical) and cleaned returns

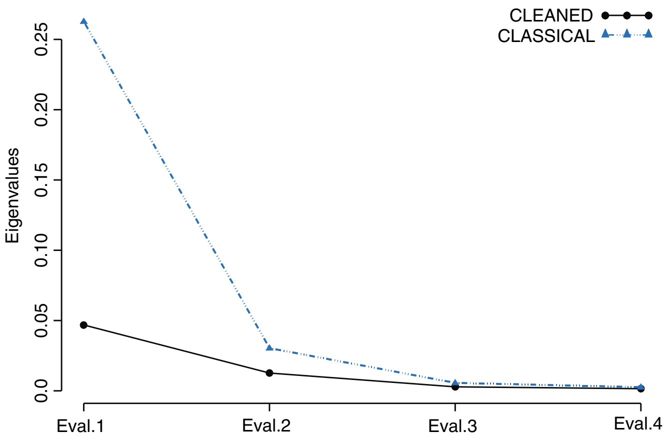


Fig. 11.5 Eigenvalues of covariance matrix estimates for original and cleaned data

While we can easily get the resulting covariance estimates based on the original and cleaned data by combining the correlation and volatility estimates, the results are not so easy to interpret. A more convenient way to assess the influence of outliers on the covariance matrix estimate is by examining plots of the eigenvalues of the sample covariance matrix for the original and cleaned data (a so-called "scree" plot) as shown in Fig. 11.5.

It is evident that the primary influence of the outliers in this case is substantial and is primarily on the largest eigenvalue.

11.2.3 Influence Functions

In order to obtain general representations of the influence of returns outliers (and inliers) on estimates of returns means, correlations, covariances, and efficient frontiers, we make use of *empirical* and *theoretical influence functions* borrowed from the statistical literature. These are very useful tools for comparing classical normal distribution maximum likelihood estimates (or small variants thereof) with robust alternatives. For a thorough discussion in the context of portfolio construction, see Scherer and Martin (2005). Here we give a very brief treatment of influence functions for portfolio construction.

Empirical influence functions (EIFs) are defined in the following in a very intuitive manner. Let $\mathbf{R}$ be the $T \times N$ matrix whose tth row contains the returns vector $\mathbf{r}_t$ and $\hat{\boldsymbol{\theta}}(\mathbf{R})$ be a vector- or matrix-valued estimated of an unknown parameter $\boldsymbol{\theta}$. Let $\mathbf{r}$ be a new vector of returns with corresponding estimate $\hat{\boldsymbol{\theta}}(\mathbf{r}, \mathbf{R})$ based on $\mathbf{r}$ and $\mathbf{R}$. Then an EIF may be defined as

$$\mathrm{EIF}(\mathbf{r}; \hat{\boldsymbol{\theta}}, \mathbf{R}) = (T + 1)(\hat{\boldsymbol{\theta}}(\mathbf{r}, \mathbf{R}) - \hat{\boldsymbol{\theta}}(\mathbf{R})),$$

where dividing the difference in estimate values by $1/(T + 1)$ provides a normalization across sample sizes T. It is easy to check that for the sample mean estimate of mean returns, the EIF has the following unbounded form in $\mathbf{r}$:

$$\mathrm{EIF}(\mathbf{r}, \bar{\mathbf{r}}, \mathbf{R}) = \mathbf{r} - \bar{\mathbf{r}}_t.$$

This means that a single outlier $\mathbf{r}$ can distort the mean returns estimate by an arbitrarily large amount.

An advantage of the use of EIFs is that because of their empirical nature they can be computed for virtually any estimator – no matter how complex – by using a suitable representative choice of the returns matrix $\mathbf{R}$. However, for estimates of portfolio construction parameters that have convenient forms in large samples, it is more convenient to use theoretical influence functions that are motivated by the limiting form of the EIF as the sample size T tends to infinity. In this case, the estimate $\hat{\boldsymbol{\theta}}(\mathbf{R})$ is replaced by its asymptotic value as a functional $\boldsymbol{\theta}(F)$ of the distribution (function) F of the returns $\mathbf{r}_t$ and the EIF difference quotient above is replaced by a directional (Gateaux) derivative $\mathrm{IF}(\mathbf{r}; \boldsymbol{\theta}(F), F)$ of $\boldsymbol{\theta}(F)$ in the "direction" defined by the mixture distribution $F_\gamma = (1 - \gamma)F + \gamma \cdot \delta_{\mathbf{r}}$, where $\delta_{\mathbf{r}}$ is a point mass distribution located at the returns data value $\mathbf{r}$. It turns out that the influence function provides an approximate measure of the maximum bias caused by $\mathbf{r}$ for small values of γ, i.e., for small fractions of outliers all located at $\mathbf{r}$. For a simple introduction, see Section 6.11 of Scherer and Martin (2005) and for a more in-depth treatment, see Hampel et al. (1986) and Maronna et al. (2006).

The sample mean and sample covariance estimates have the large-sample functional forms

$$\boldsymbol{\mu}(F) = \int \mathbf{r} \mathrm{d}F(\mathbf{r}),$$
$$\boldsymbol{\Omega}(F) = \int (\mathbf{r} - \boldsymbol{\mu}(F))(\mathbf{r} - \boldsymbol{\mu}(F))' \mathrm{d}F(\mathbf{r}),$$

and their influence functions turn out to be (see the references cited above):

$$\mathrm{IF}(\mathbf{r};\boldsymbol{\mu}(F),F) = \mathbf{r} - \boldsymbol{\mu},$$
$$\mathrm{IF}(\mathbf{r};\boldsymbol{\Omega}(F),F) = (\mathbf{r} - \boldsymbol{\mu})(\mathbf{r} - \boldsymbol{\mu})' - \boldsymbol{\Omega},$$

where we have simplified the notation on the right-hand-side of the above equations by using $\boldsymbol{\mu} = \boldsymbol{\mu}(F)$ and $\boldsymbol{\Omega} = \boldsymbol{\Omega}(F)$. These two influence functions lack a type of robustness in that they are both unbounded, the first in a linear sense and the second in an even worse quadratic manner.

Note that in the case of a single asset, the covariance matrix collapses to the variance functional $\sigma^2(F)$ and the influence function is

$$\mathrm{IF}(r,\sigma^2(F),F) = (r-\mu)^2 - \sigma^2.$$

Note also that since the influence function is a derivative, applying the chain rule yields the influence function for the sample volatility estimate

$$\mathrm{IF}(r,\sigma(F),F) = \frac{(r-\mu)^2 - \sigma^2}{2\sigma^2},$$

which perhaps surprisingly is still quadratic in r.

Any return value r in the interval $\mu \pm \sigma$ has a negative influence with σ^2 as the maximum negative influence. Negative influence corresponds to negative bias in a variance or volatility estimate, and it is natural to use the term *inlier* for any value of r that has negative influence. The practical relevance of this point is the obvious one that a thinly traded asset that produces a zero returns value will cause negative bias in volatility and variance estimates.

11.3 Basic Robustness Concepts

This section is intended to give the reader an understanding of some of the key concepts of robustness in a brief and reasonably informal manner. For a more in-depth treatment, see Maronna et al. (2006) and the references therein, as well as the earlier classic books by Huber (1981) and Hampel et al. (1986).

11.3.1 Data-Oriented Robustness

As is evident from the comment in Sect. 11.1, statistical robustness is primarily focused on methods that are not much influenced by outliers. More generally, an estimator should not be changed very much by "large changes in a small fraction of the data" or "small changes in most of the data." The first requirement speaks to the need for insensitivity to outliers or unbounded moments beyond the first moment, while the second refers to insensitivity, for example, to rounding errors in all the data. A robust estimator that meets these requirements should deliver the following desirable behavior in the portfolio context:

- A good model fit to the bulk of the returns
- Reliable detection of multidimensional returns outliers
- Diagnostics on risk and performance impact of outliers
- Increased stability of MVO optimal portfolios
- Stable prediction of typical future returns

Examples in the reminder of the paper will clearly illustrate the achievement of the first three properties. The fourth and fifth properties could be revealed by the studies of robust factor models for returns prediction and out-of-sample back tests of robust MVO portfolios, both of which remain topics for future work.

11.3.2 Robustness Theory

In addition to the above intuitive data-oriented concepts, there fortunately exist the following theoretical concepts of robust estimation, among others:

- Variance efficiency robustness
- Bounded influence robustness
- Bias robustness

The above concepts are construed within the framework of "doing well near a normal distribution," i.e., the central bulk of the data is well modeled by a normal distribution but the data in the tails follow a more fat-tailed and possibly skewed distribution (Huber 1981).

11.3.2.1 Variance Efficiency Robustness

This old approach to robustness attempts to find estimators, e.g., of mean returns, correlations, and covariances, that have close to the minimum attainable variance across a finite set of distributions containing the normal distribution and an "approximately spanning" set of fat-tailed distributions. This line of research was initiated by Tukey (1960) and resulted in the so-called "Princeton Study" results reported in Andrews et al. (1972). These works no doubt provided stimulation for the seminal work on robust estimates of location by Huber (1964) that obtained min–max variance robust estimates of location (the mean in finite mean situations) across an infinite family of mixture distributions with central component a normal distribution. We note that little attention was given in the above works to skewed distribution deviations from normality and the resulting bias problem. A small section of Huber

(1964) focuses on the mixture distributions with symmetric central component and skewed mixing component, and shows that the sample median minimizes the maximum bias. It is fair to say that this important result is not very well known, even to professional statisticians, and deserves to be better known.

11.3.2.2 Bounded Influence Robustness

The approach of constructing robust estimators with bounded influence functions is an intuitively appealing one that was pioneered by Hampel (1974), who constructed optimal bounded influence estimators that control bias due to a small fraction γ of outliers while achieving high normal distribution efficiency. This influence function approach is the focus of Hampel et al. (1986).[2] Choosing robust estimators with attractive influence function properties is indeed a viable approach that we emphasize in Sect. 11.4.

11.3.2.3 Bias Robustness

It is a basic statistical fact that the mean-squared-error of an estimator is the sum of the variance and the squared bias, with the variance typically going to zero at a rate of n^{-1}. However, outliers can cause bias that does not disappear in large samples, and in fact squared bias can dominate variance in surprisingly small samples. This suggests focusing on robustness with respect to bias, i.e., controlling the maximum bias due to outliers. The fact that the influence function choice only gives control of bias for small fractions of outliers left open the problem of bias robust estimators that minimize the maximum bias due to outliers for any choice of outlier fraction $\gamma < 1/2$. Solutions to this problem are available for estimating scale (Martin and Zamar 1993a) and regression (Martin et al. 1989). However, sometimes a solution that focuses only on controlling maximum bias can have bad variance properties. This, for example, is the case with the least-median-of-squares regression estimate, which is known to be a very good approximation to a min–max bias robust regression estimate (Martin et al. 1989). Subsequently Martin and Zamar (1993b) obtained min–max bias robust location estimates subject to a constraint on efficiency at the normal model. Then Svarc et al. (2002) did likewise for the much more important problem of regression. When applied to the estimation of mean returns alone, the estimator found by the latter authors has the desirable influence function shown in the lower right of Fig. 11.6. It is not often that such theoretical bias robust solutions may be obtained, and indeed no such solution is known in the

[2] Hampel's optimality problem is to minimize the estimator variance at the normal (or other nominal) distribution subject to a bound on the maximum of the influence function, i.e., a bound on the so-called *gross-error sensitivity* (GES). The problem has interesting solutions (see Hampel et al. 1986).

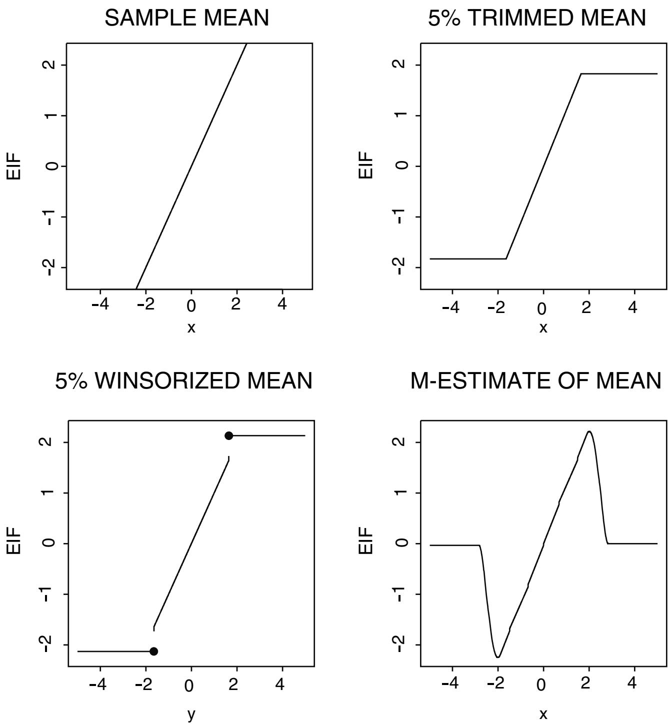


Fig. 11.6 Empirical influence functions of mean returns estimates

case of estimating correlations and covariances in general dimension. Still it is a desirable goal to choose robust estimators that have bounded influence and good control of bias due to outliers.[3]

11.4 Robust Means, Covariances, and Correlations

There are a variety of ways of constructing robust estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ with the robustness properties of Sect. 11.3. As the preceding discussion suggests, we will focus for purposes of this paper on a few possibilities for robust estimates of mean returns taken one asset at a time, and a single reasonable attractive robust covariance and correlation matrix estimate. The discussion in Sect. 11.4.1 is intended to

[3] There are some subtleties in this research arena. For example, the min–max bias robust estimator of Svarc et al. (2002) version of the MM-estimate of Yohai (1987) has an unbounded influence function. But while its maximum bias is strictly worse than that of some bounded influence regression estimates for small γ, it has small maximum bias for a wide range of $\gamma < 1/2$. See also Yohai and Zamar (1997).

be informative with respect to the influence functions of three alternative choices, with a view toward rationalizing the choice of robust covariance matrix estimate in Sect. 11.4.2.

11.4.1 Trimmed Means, Winsorized Means, and M-estimates

Here we describe the two most frequently used simple robust estimates of mean returns that are in common use. Let $r_{1:n} \le r_{2:n} \le \cdots \le r_{n:n}$ be the ordered returns. An α*-trimmed mean* is computed by discarding a fraction α of the largest and smallest ordered returns and computing the sample mean of the remaining returns. For example, when $n = 60$, a 5% trimmed mean is computed using $\alpha = 0.05$ as

$$\hat{\mu}_{\text{trim, 5\%}} = \frac{1}{54}\sum_{i=4}^{57} r_{i:60}.$$

An α*-Winsorized mean* is computed by "pulling" a fraction α of the largest and smallest order statistic values to the next largest and next smallest values, respectively, and computing the sample mean of the remaining data points. So a 5% Winsorized mean is computed as:

$$\hat{\mu}_{\text{winsor, 5\%}} = \frac{1}{60}\left(4r_{4:60} + 4r_{57:60} + \sum_{i=5}^{56} r_{i:60}\right).$$

Trimmed and Winsorized means are evidently the most often used robust estimators in portfolio construction applications, with α in the range of 1–10%.

Figure 11.6 displays the EIFs of the sample mean estimate of returns along with that of a 5% alpha-trimmed mean, a 5% Winsorized mean, and a so-called M-estimate based on a redescending "psi-function."[4] They are computed "at the normal distribution" by virtue of using 60 quantiles of equally spaced probabilities between zero and one as the single column returns configuration matrix $\mathbf{R}$ of Sect.11.2.3. These EIFs are very close approximations to their theoretical influence functions computed as described in Sect.11.2.3. The main points of these influence functions are as follows:

- The sample mean has an unbounded influence function with slope one.
- The trimmed mean has a bounded influence function that is linear in a central region. Note that while trimmed returns are deleted in computing the resulting trimmed mean, the returns that are trimmed none-the-less have nonzero influence on the estimate.

[4] An M-estimate of the mean returns of a single asset is a solution of a nonlinear equation of the form $\sum_{t=1}^{T} \psi((r_t - \hat{\mu})/\hat{s}) = 0$ where $\hat{s}$ is a robust scale estimate. For details see Maronna et al. (2006) and for applications to returns see Scherer and Martin (2005).

- The Winsorized mean behaves similarly to the trimmed mean except that it somewhat surprisingly has a discontinuity at each of the upper and lower Winsorizing points. The extreme returns still have nonzero influence.
- The M-estimate has a redescending shape that is linear in the central region and redescends quickly but smoothly to zero. This feature is highly desirable because, unlike the behavior of the trimmed and Winsorized means, all returns outliers that are sufficiently large in absolute value have zero influence on the estimate.

11.4.2 The Minimum Covariance Determinant Estimate

The minimum covariance determinant (MCD) estimate was originally proposed by Rousseeuw (1985). See also Rousseeuw and Leroy (1987). The "ideal" MCD estimate with tuning parameter γ finds a subset of the returns $\mathbf{r}_t,\ t = 1,\ldots,T$ containing a fraction $1-\gamma$ of the T observations that has the minimum value of the determinant of the classical covariance estimate. For the examples that follow, we used the *fast MCD* algorithm due to Rousseeuw and van Driessen (1999) that finds a good approximate solution $\hat{\boldsymbol{\Omega}}$ to the ideal MCD estimate, as implemented in S-PLUS (2002). A robust correlation matrix is obtained by dividing the elements of the robust covariance matrix by the square root of the diagonal elements in the usual manner. The fraction $1-\gamma$ of observations that yield the MCD estimate provides a robust mean returns estimate $\hat{\boldsymbol{\mu}}$.

It can be shown that the influence function of the MCD estimate behaves like that of the classical sample covariance estimate given in Sect. 11.2.3 in a bounded central elliptical region where it exhibits bounded quadratic behavior, but has value zero outside that region (Croux and Haesbroeck 1999). The latter aspect corresponds to the fact that a fraction γ of the returns $\mathbf{r}_t$ are ignored at the MCD solution point. Thus outlier values $\mathbf{r}_t$ have at most a bounded influence on the MCD covariance estimate and associated covariance and mean returns estimates.

11.4.3 Application to Four Portfolios

We first show application of the robust covariance and resulting correlation estimates for the small-cap portfolio of Sect. 11.2.1. The results shown in Fig. 11.7 should be compared with those in Fig. 11.4. We see that the classical correlation estimates are the same in both figures and that the robust correlations in Fig. 11.7 are similar to the classical correlations based on the cleaned data in Fig. 11.4, with the robust correlations in Fig. 11.7 being generally somewhat larger than the classical correlations for the cleaned data in Fig. 11.4. We note that the outliers influence on the correlation pairs (REGN, RVT) and (REGN, CHD) results in a very optimistic view of their diversification opportunities, grossly so in the latter case where the difference in correlations is 0.49.

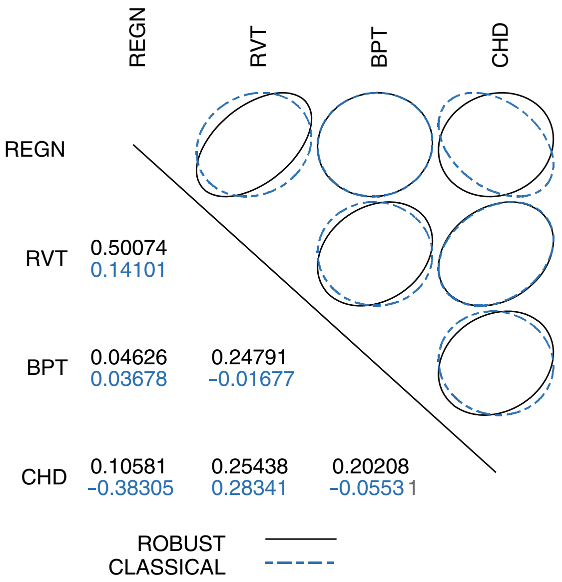


Fig. 11.7 Robust and classical correlations for small-cap portfolios

Figures 11.4 and 11.7 show that robust correlation estimation does automatically what careful manual outlier cleaning accomplishes. While one can implement one-dimensional outlier cleaning with one of the estimates of Sect. 11.4.1, this will not reliably identify influential multidimensional outliers, as is shown in Sect. 11.6.

Figures 11.8–11.10. follow display robust and classical correlations for the hedge funds, ETFs, and commodities portfolios.

We note that outlier influence on the classical correlation estimates exists for each of the three portfolios in Figs. 11.8–11.10. Two pairs in Fig. 11.8 have differences of 0.47 and 0.51 between the classical and robust estimates, two pairs in Fig. 11.9 have differences of 0.36 and 0.41, and two in Fig. 11.10 have differences of 0.28. It is to be noted that often the robust correlations are larger than the classical ones, indicating that the classical estimates are too optimistic about diversification opportunities. However, the opposite sometimes occurs, i.e., the classical estimate may be too pessimistic about diversification opportunities. For example, in the case of the hedge funds portfolio of Fig. 11.8, the correlation estimate for the pair (F5, F8) has a classical estimate of −0.19 and a robust estimate of 0.32, while for the pair (F2, F4) the classical estimate is 0.45 compared with the robust estimate of −0.11.

Figure 11.11 displays the robust and classical eigenvalue patterns for the three portfolios. In the case of the hedge funds portfolio, all the eigenvalues are larger for the robust covariance matrix estimate than the classical estimate, while the gap between the first two eigenvalues of the robust covariance are smaller than for the classical covariance. Thus the robust covariance "shape" is a little "fatter," less "elongated" ellipse in eight-dimensional space, which evidently implies a bit

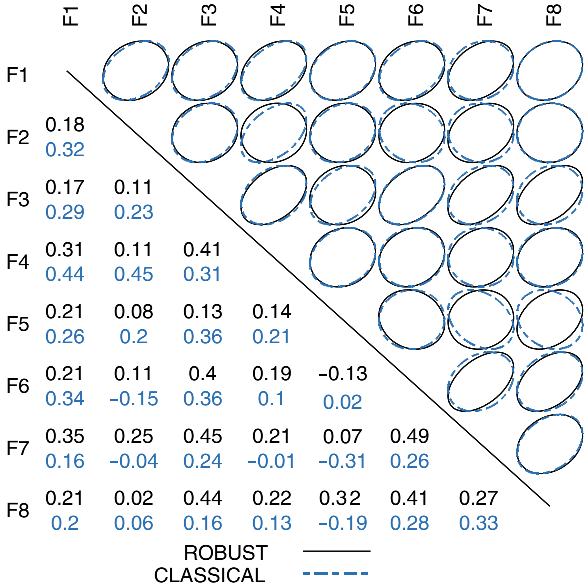


Fig. 11.8 Robust and classical correlations for hedge funds portfolio

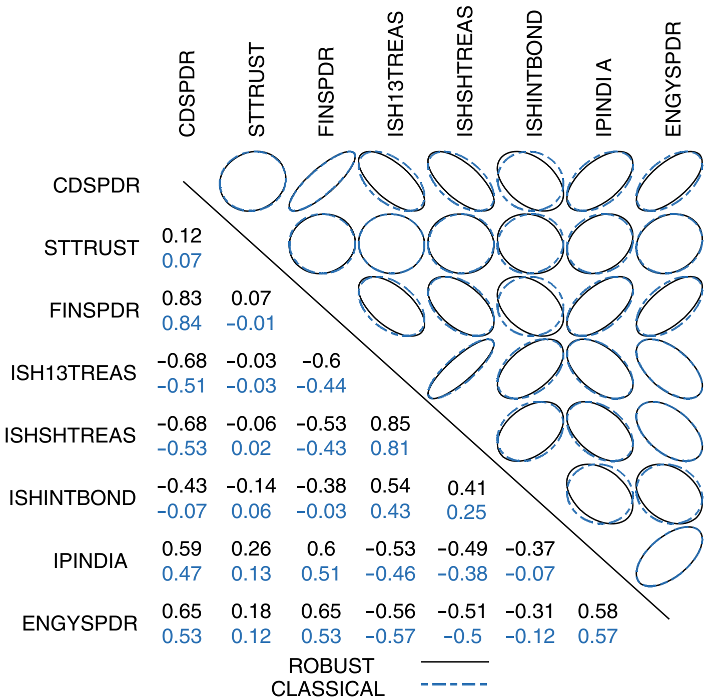


Fig. 11.9 Robust and classical correlations for ETF portfolio

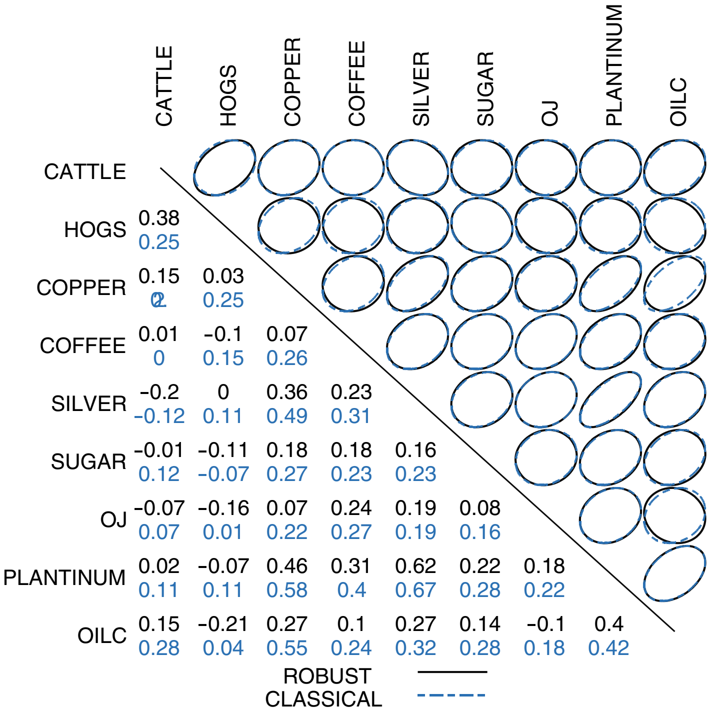


Fig. 11.10 Robust and classical correlations for commodities portfolio

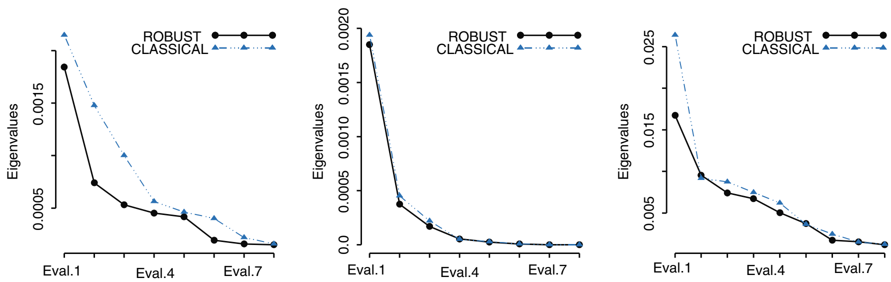


Fig. 11.11 Robust and classical eigenvalue plots for portfolios of Figs. 11.8–11.10

more overall diversification opportunity in the portfolio than the classical covariance implies. The shape and implications are somewhat the opposite for the commodities portfolio, and there is virtually no difference in the eigenvalues for the ETFs portfolio.

11.5 Robust MVO Efficient Frontiers

We now apply robust mean and covariance estimates to the construction of robust mean–variance (MVO) portfolios by simply substituting the robust estimates for the classical estimates, either in a quadratic utility function or in minimizing risk for target levels of return. In doing so, we restrict attention to long-only fully invested portfolios with no additional constraints and display robust and classical MVO efficient frontiers for each of the four example portfolios. The efficient frontier plots are displayed with monthly returns and volatility coordinates in decimal form and risk-free rate of 0.003 monthly, except for the ETF portfolio which is based on daily returns and volatility and a risk-free rate of zero. In all four examples, there are substantial differences in mean return and risk across the efficient frontiers that would be of concern to the professional portfolio manager.

11.5.1 Small-Cap Efficient Frontiers

Figure 11.12 shows the robust and classical efficient frontiers for a toy small-cap portfolio of the four stocks REGN, RVT, BPT, and CHD. Clearly the huge outlier in the REGN returns totally distorts the classical efficient frontier. The robust efficient frontier covers only a small range of the return and risk values and at the minimum risk position has allocations of approximately 70% CHD, 26% RVT, 4% BPT, and 0% REGN.

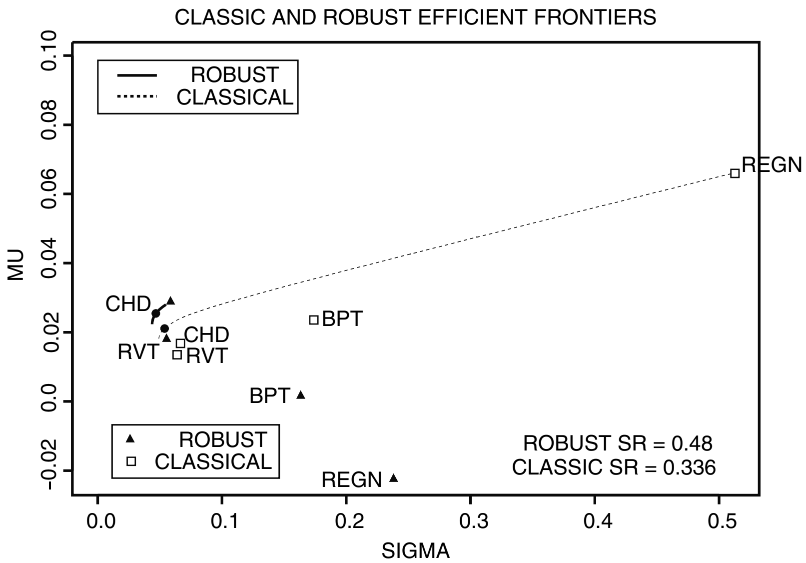


Fig. 11.12 Small-cap portfolio efficient frontiers

11.5.2 Hedge Fund Portfolios

In order to reveal the different kinds of configurations of robust and classical efficient frontiers, we computed them for four portfolios based on three different subsets of size four of the eight hedge funds whose EDA plots of monthly returns are provided in the Appendix. The results are displayed in Fig. 11.13. The top two plots in Fig. 11.13 are for the same portfolios consisting of F1, F2, F3, F4 except the first is based on the entire 5-year period starting from 10/31/99 and ending in 9/30/04, which includes the peak and collapse of the dotcom bubble, and the second is based on the 4-year period starting from 10/31/00 which excludes the initial collapse of the dotcom bubble. The bottom two plots are for the two portfolios consisting of F1, F2, F7, F8 and F1, F4, F5, F6, both for the 4-year period starting from 10/31/00.

For the portfolio F1, F2, F3, F4, we see in the top left plot that when the dotcom peak and collapse is included there is a considerable difference between the two efficient frontiers and they actually cross each other at a point near the minimum variance classical MVO portfolio. In top right plot, which excludes the dotcom peak and collapse period from 10/31/99 to 10/30/00, the two efficient frontiers are virtually identical. This indicates that all the influential outliers occurred during that period. Perusal of the time series plots of these four hedge funds in Appendix makes it abundantly clear that there is a regime shift once the dotcom collapse is ended (perhaps somewhat later than 10/30/00). The careful portfolio manager either would not include the first few years data in an analysis relevant for Q4 of 2004 or

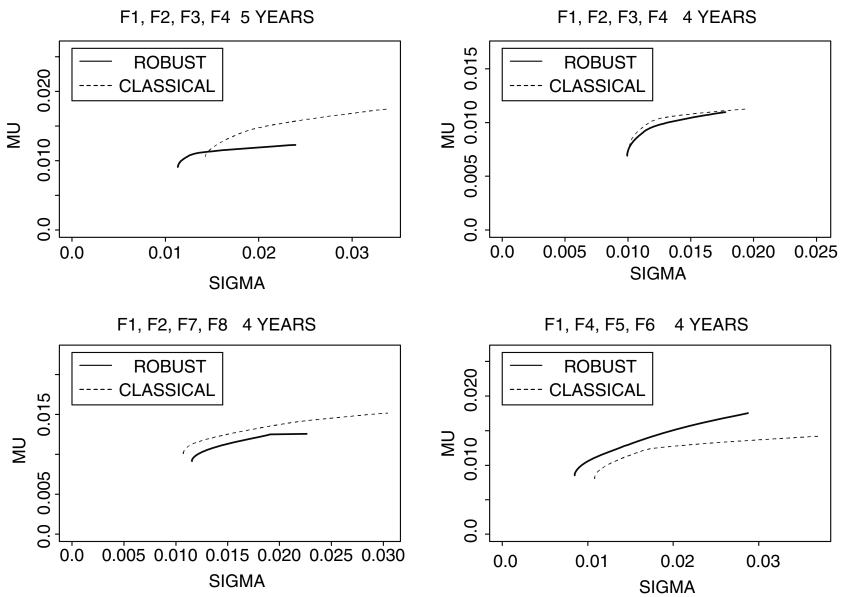


Fig. 11.13 Efficient frontiers for three hedge fund portfolios and two time periods

would use robust methods to construct a regime shift model using a hidden Markov process switching model based on an autoregressive-moving average model or a vector error correction model (VECM).

If the manager's choice were to exclude data from her sample, taking into account the higher returns and volatility associated with the period 9/31/99 to 10/30/00, it is clear that the robust efficient frontier with its lower return and risk profile is the more appropriate one to use in a forward looking view.

This example illustrates what is often the case with a good robust method: it comes to a conclusion similar to that which a portfolio manager would, provided she had done a sufficiently careful analysis. Unfortunately, time pressure or the subtlety of outlier configurations may result in the portfolio manager not discovering the influential data in the returns history used for portfolio construction; however, the robust method will usually do so and make the needed adjustments *automatically* in the estimation process.

The two portfolios in the bottom two plots illustrate two different configurations that can occur besides the crossing pattern in the top left plot, namely, that a classical MVO efficient frontier can dominate a robust MVO efficient frontier and conversely (though the latter seems to occur more frequently, as in Figs. 11.14 and 11.15). In Sect. 11.7, we discuss the implications of these different efficient frontier patterns that can occur in the context of the risk analysis diagnostics they provide.

11.5.3 ETFs Portfolio

See Fig. 11.14.

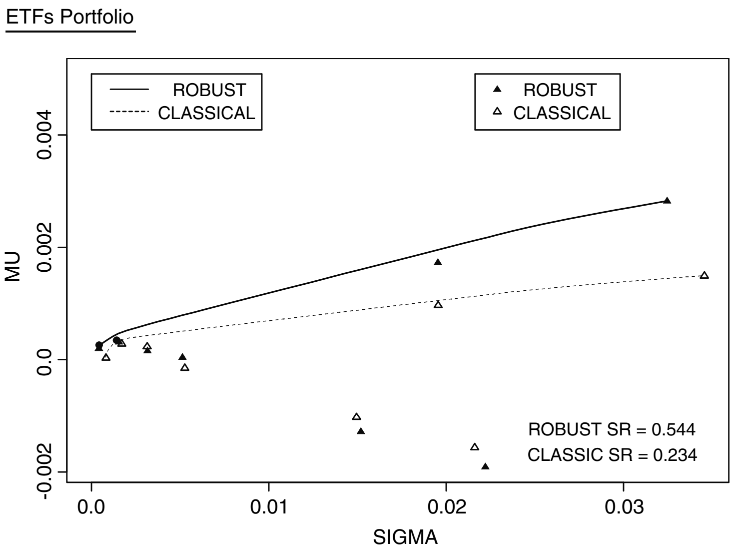


Fig. 11.14 Efficient frontiers for the ETFs portfolio (rf = 0%)

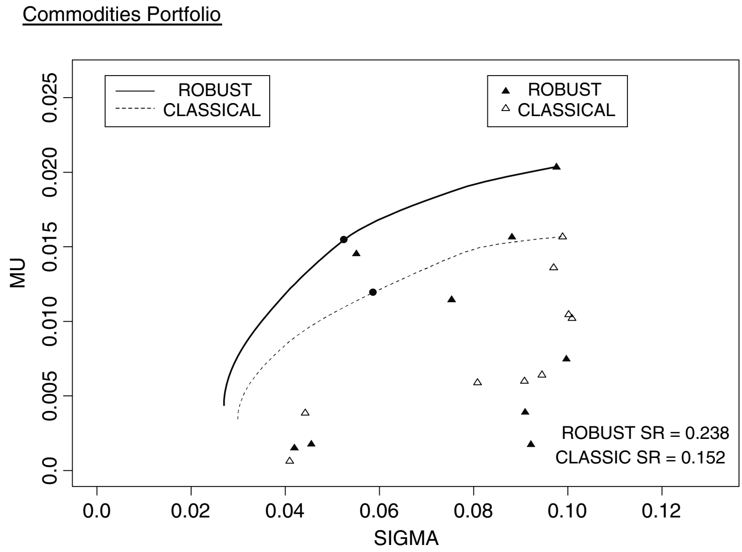


Fig. 11.15 Efficient frontiers for commodities portfolio

11.5.4 Commodities Portfolio

See Fig. 11.15.

11.6 Robust Returns Outliers Detection

Among those (relatively few) portfolio managers who are aware that outliers can have an adverse influence on their portfolio management practice, most seem to think that one-dimensional outlier treatment such as one-dimensional trimming or Winsorizing treatment will suffice. Such treatment is taken to mean either deleting or shrinking a fraction of the extreme values as defined in the trimmed or Winsorized means in Sect. 11.4.1, or replacing extreme values of the ordered returns by a robust mean estimate. Indeed, such one-dimensional treatment is much better than doing nothing about outliers and sometimes it will suffice. However, it sometimes happens that multidimensional returns outliers exist that are not outliers in any single one-dimensional set of returns; such multidimensional outliers can have a severe influence on covariance matrix estimates and ensuing optimal portfolio calculations. A particularly striking example of this is provided in Sect. 11.6.2. But first we show in Sect. 11.6.1 that robust covariance estimation is needed for reliable multidimensional returns outlier detection-based automatic alerts to the portfolio manager. In Sect. 6.3, we discuss the need for robustness in testing for nonnormality in both one-dimensional and multidimensional tests for nonnormality.

11.6.1 Detection of Multidimensional Returns Outliers

A natural way to detect multidimensional returns outliers is to use the distances $d(\mathbf{r}_t)$ defined by quadratic forms

$$d^2(\mathbf{r}_t) = (\mathbf{r}_t - \hat{\boldsymbol{\mu}})'\hat{\mathbf{C}}^{-1}(\mathbf{r}_t - \hat{\boldsymbol{\mu}}), \quad t = 1, \ldots, T,$$

where $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{C}}$ are good estimates of the true mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$ of the N-dimensional return vector $\mathbf{r}_t$. It may be shown that the geometric action of the above distances is to rotate a central elliptical scatter to diagonal coordinates and then scale the coordinates to obtain a central sphere of data, to which ordinary Euclidean distance is then applied (see for example Scherer and Martin 2005). Under multivariate normality and with the estimates replaced by their true values, the above squared distance has a chi-squared distribution χ^2_N with N degrees of freedom. The distribution will be approximately χ^2_N when the estimates are close to their true values, and one can compare the $d(\mathbf{r}_t)$ with an upper tail percent point of the χ^2_N distribution, e.g., the upper 1% point.

However, we know from the results of Sect. 11.4 that the classical covariance estimates may be distorted by outliers and may thus be poor estimates of the covariance structure of the bulk of the returns data. So we construct robust distances by using robust estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{C}}$ in the above distances in order to obtain reliable outlier detection. The results in Figs. 11.16–11.19 for our four example portfolios speak for themselves. For the first three of these figures we used $\gamma = 0.05$, resulting in the robust MCD covariance estimate rejecting three returns vectors for the

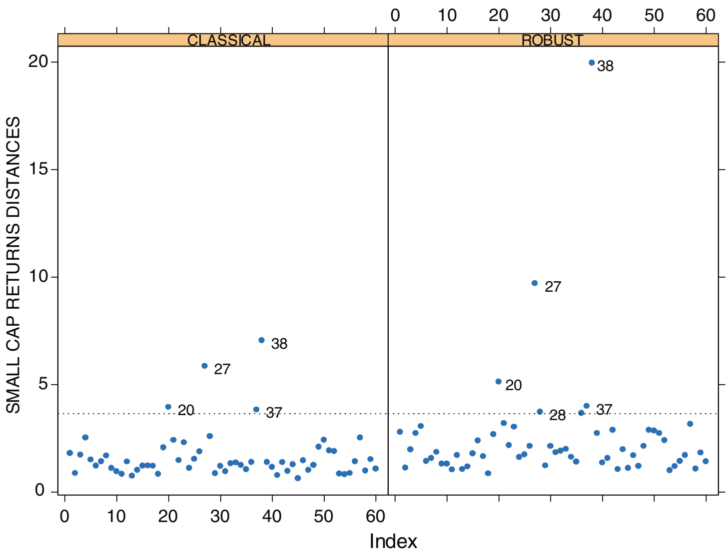


Fig. 11.16 Returns distances for small-caps

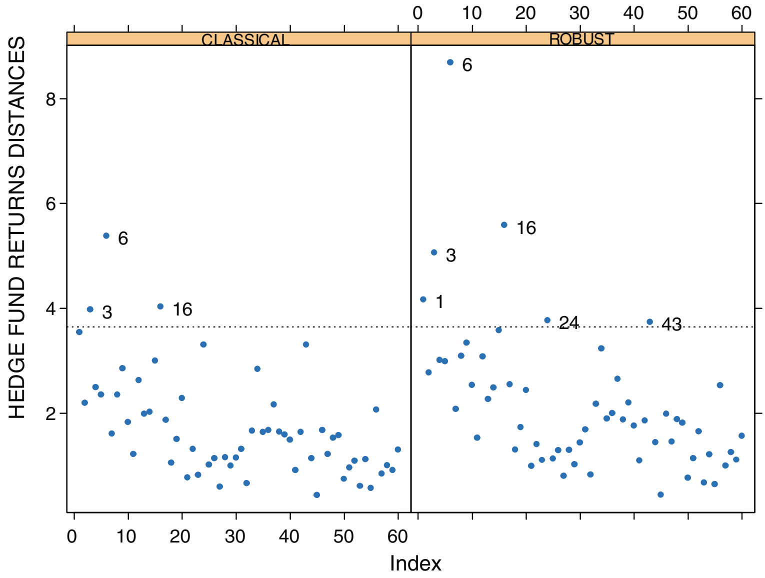


Fig. 11.17 Returns distances for hedge funds

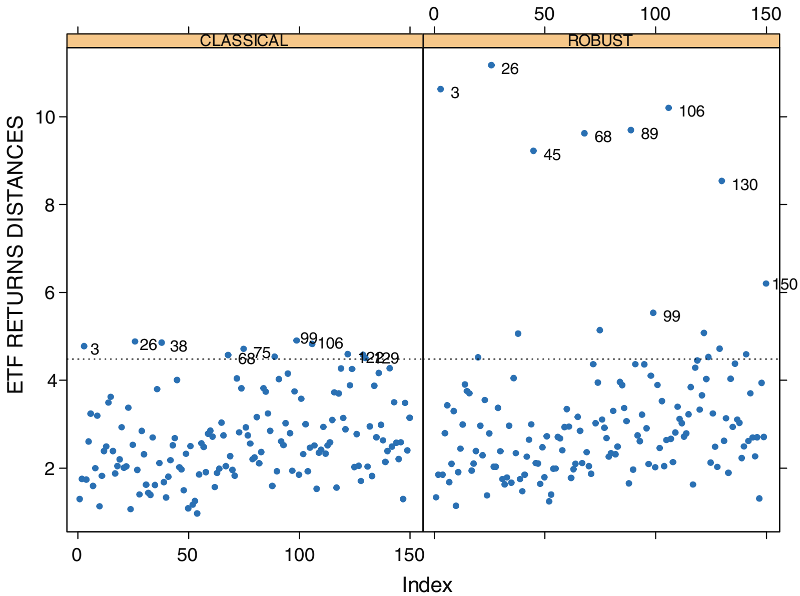


Fig. 11.18 Returns distances for ETFs

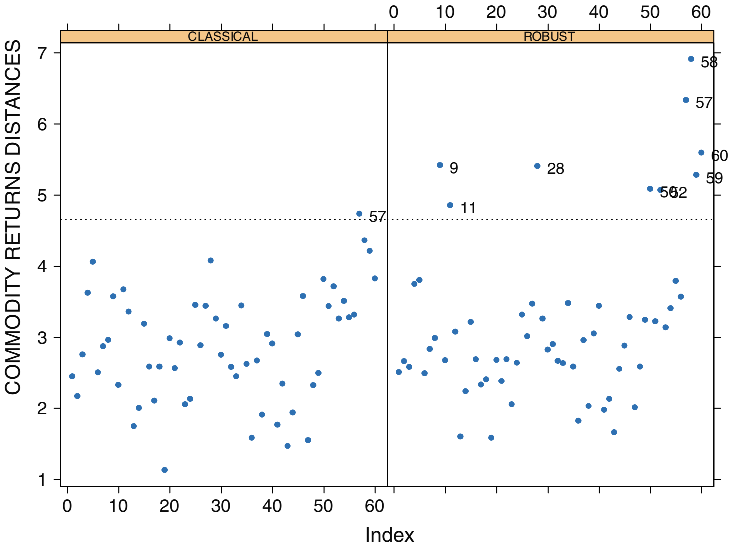


Fig. 11.19 Returns distances for commodities

small-cap and hedge fund portfolios, each of which had 60 observations, and rejecting seven returns vectors for the ETFs portfolio, which had 150 observations. We used $\gamma = 0.15$ for the commodities portfolio resulting in rejection of nine observations since this portfolio had a total of 60 observations. (We also use this value for the ellipses and eigenvalue plots presented earlier.) We comment further on this choice subsequently.

For the small-caps portfolio, the classical distances in Fig. 11.16 detect four outlier returns of dimension four, two of them marginally, while robust distances detect five outliers, two of them marginally. The relative improvement in outlier detection when using the robust distances is rather similar for the hedge funds in Fig. 11.17.

The improvement of the robust method over the classical method is more dramatic for the ETFs in Fig. 11.18, where the classical method, though formally detecting a number of outliers, does not reveal meaningful separation between them and the bulk of the data. The robust distances, by way of contrast, clearly separate most of the outliers from the bulk of the data.

The improvement of the robust detection method over the classical method is equally dramatic for the commodities in Fig. 11.19.

The motivation for experimenting with the parameter γ, i.e., the fraction of returns vectors rejected in the MCD, and finally using $\gamma = 0.15$ for the commodities portfolio is that perusal of the time series plots of the commodities returns reveals considerable volatility clustering and regime shifts, giving rise to the suspicion that these returns contain a large fraction of multivariate outliers. Indeed using $\gamma = 0.05$ for the commodities returns resulted in detection of only the last three outliers. While we have found that often a relatively small value such as $\gamma = 0.05$ works well, this of course begs the question of how to choose γ since the portfolio manager would not know in advance how many outliers there are. We conjecture that an effective approach would be to choose γ to maximize a measure of nonnormality of the robust distances when transformed from a nominal chi-squared distribution to a normal distribution, e.g., maximizing kurtosis or the Jarque–Bera statistic which combines kurtosis and skewness. This issue remains to be studied.

11.6.2 Fundamental Factor Model Exposures Outliers

We now focus on one aspect of fitting fundamental factor models, namely the need to deal with outliers in the exposures/loadings matrix, that is well recognized by practitioners. It is common practice on the part of equity portfolio managers and by commercial equity portfolio management software providers to trim or Winsorize the exposures one at a time prior to fitting a fundamental factor model, typically by a weighted least squares method. Here we show quite clearly that such one-dimensional outlier treatment of the exposures does not suffice to eliminate multidimensional outliers that have the potential to adversely influence least squares fits of fundamental factor models.

The data for our experiment consists of the four-dimensional vectors of book-to-market, earnings-to-price, size (i.e., log of market capitalization in millions of dollars), and a momentum factor consisting of 12-month moving averages of returns for 1,046 equities. Figure 11.20 displays the 1,046 distances for the returns

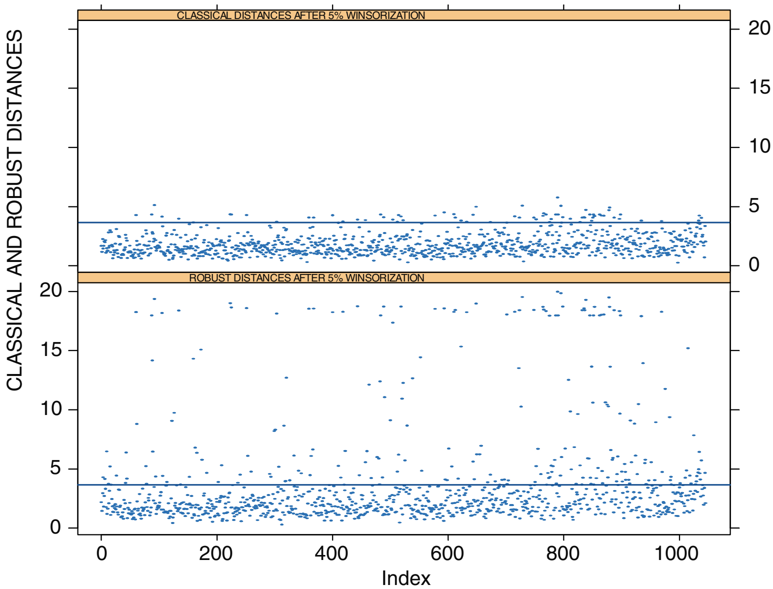


Fig. 11.20 Robust and classical distances after 5% Winsorizing of exposures

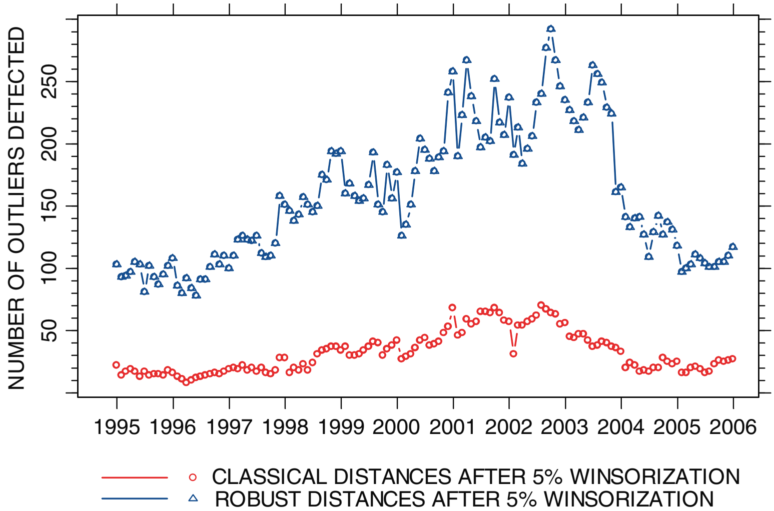


Fig. 11.21 Number of four-dimensional exposures outliers detected

for a single month on 07/31/2002, and Fig. 11.21 shows the time series of the number of outliers detected each month, for both classical and robust distances. Both figures are based on doing one-dimensional Winsorizing at 5% on each of the four risk factor exposures and then computing classical and robust distances for the four-dimensional exposures vector. Two-dimensional scatter plots show that most of the multidimensional outliers detected by the robust distance plots after Winsorizing are not even two-dimensional outliers let alone one-dimensional outliers.

Clearly one-dimensional outlier treatment in fundamental equity factor models will not suffice. The pervasiveness and size of the multidimensional outliers that occur in fundamental factor model exposures indicate that, at a minimum, bias robust regression methods will be needed to reliably fit such models. For an example of the efficacy of robust regression in estimating beta in single factor models see Martin and Simin (2003). One of us (Clark) has used normal mixture models and multidimensional outlier detection to construct value and growth indices using four-factor models with good success.

11.6.3 Other Methods of Multidimensional Outlier Detection

The problem of detecting multidimensional outliers is a fundamental and important problem in applied statistics. Though techniques such as the Mahalanobis distance have been used for some time, their frequent unreliability has led to the development of techniques which have been known in the statistical community for well over a decade now. We used the dynamical graphic technique called the “grand tour" (the grand tour is defined as a complete multidimensional look at the data via approximately all three-dimensional slices, an analysis that we completed using the Statistica software product) and cluster analysis to detect multiple outliers in our multidimensional portfolio returns.

Agglomerative complete linkage clustering, which produces a topology-preserving mapping of the multidimensional outliers onto a two-dimensional plane, gives us an easy way of visually detecting multidimensional outliers in the data. Our grand tour/cluster-based method of outlier detection not only identifies the multi-dimensional outliers, but also actually provides information about the entire outlier neighborhood via the dendogram produced by the cluster analysis. We applied our method to detect outliers from various types of multivariate datasets and showed that it can effectively be used in multidimensional outlier detection.

Figure 11.22 is a sample plot from the grand tour done on three of the four small-cap securities in our dataset. The numbered spheres are suspected outliers that will come to part of our cluster analysis. Observations 20, 27, and 38 look to be potential outliers, as they are some distance away from the bulk of the data. The remaining outliers identified by the grand tour were observations 28, 51, and 57. The dendogram derived from these six data points clearly identifies observations 38 and 27 as outliers and potentially 28 as well (Fig. 11.23).

Finally, it is interesting to note that despite the predominance of normality in the commodities data, there were five prominent outliers in the dataset when analyzed via the grand tour and cluster analysis. These outliers, if they were not removed or taken into account via robust techniques, would have clearly produced a suboptimal investment strategy.

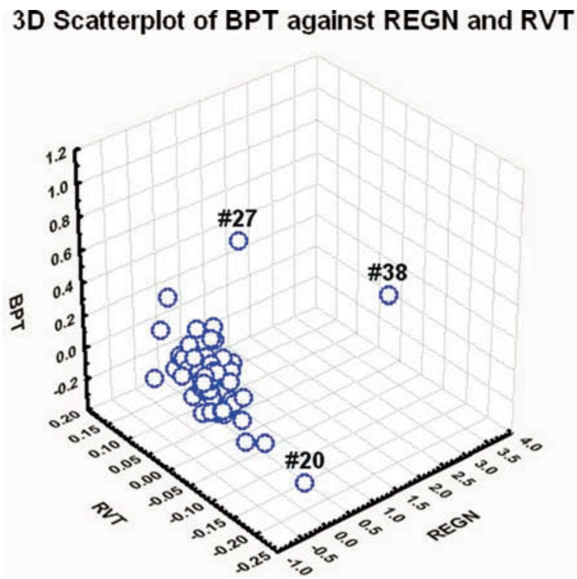


Fig. 11.22 3D scatter of small-caps returns BPT (z), REGN (x) and RVT (y)

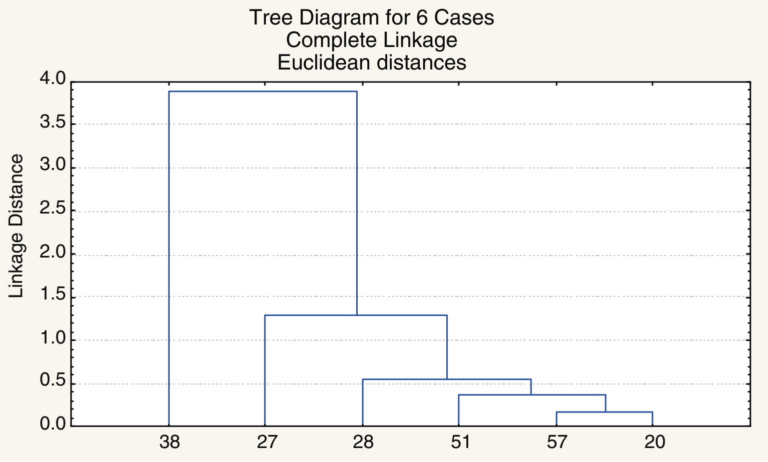


Fig. 11.23 Dendogram of small-cap outliers

11.7 Robust Portfolio Construction Diagnostics

Use of the robust methods in finance, including, but not limited to, those described here for portfolio construction, are not a be-all and end-all, and it is by no means recommended that robust MVO be used to the exclusion of other methods such as MVO or mean vs. ETL portfolio optimization (see Sect. 11.8). Instead, robust methods should be used as a complement to such methods, and above all for their corresponding comparative diagnostic value. In the words of J. W. Tukey a long time ago (circa 1979):

> "... just which robust/resistant methods you use is *not* important – what is important is that you use *some*. It is perfectly proper to use both classical and robust/resistant methods routinely, and only worry when they differ enough to matter. *But* when they differ, you should think *hard*."

The visual displays of classical and robust correlations, efficient frontiers, and distance plots as in Sects. 11.5 and 11.6 are highly effective visual alerts for making quick assessments of whether the two methods differ enough to motivate hard thinking, e.g., about the times of occurrences and possible asset conditions or financial markets or sectors causes of the outlier returns. In some cases, e.g., when there is a single highly influential outlier, when outliers represent a small fraction of the returns sufficiently far in the past, or when outliers represent identifiable data errors, the portfolio manager may decide to use the robust MVO. In other cases, e.g., when the outliers are relatively recent in time and possibly represent transition to a new regime, the portfolio manager may decide to stay with the classical MVO. In any event, the calculation of robust correlations and efficient frontiers will alert the manager to possible problems of which she would otherwise be completely unaware.

Since risk is central to evaluating portfolio performance, it is natural to focus on the diagnostic use of robust estimates in risk calculations. Since robust methods down-weight or reject outliers, one's first reaction may be that robust estimates will lead to smaller risk estimates than those obtained with classical estimates in the presence of one or more outliers, in which case the difference between the latter and the former will be positive. That is certainly the case with the sample standard deviation (volatility) as risk measure. However, this is not always the case, i.e., the difference between classical and robust risk estimate may be negative as well as positive.

For example, in Fig. 11.13, you see three distinct behaviors in the upper left, lower left, and lower right figures where the efficient frontiers differ: (a) there is a large region where the robust efficient frontier risk is larger than that of the classical efficient frontier and a small region where the opposite is true, (b) the robust efficient frontier has uniformly higher risk, and (c) the robust efficient frontier has uniformly smaller risk. The reason that case (b) and the small region in case (a) can occur is because the efficient frontier depends on means and covariances in a complex manner, e.g., an outlier can cause some correlations to be small thereby reflecting a reduced risk through diversification without a sufficiently offsetting effect in mean and variance estimates.

11.8 Fat-Tailed Distributions vs. Robust Methods

When a portfolio to be optimized is based on sufficient history of asset returns that exhibit fat-tailed and skewed distribution behavior, a highly attractive portfolio optimization approach consists of the following two ingredients: (a) model the marginal distributions of each asset as well as the cross-sectional dependency structure, and (b) use the resulting model with an appropriate down-side risk measure in a "mean-risk" alternative to MVO. An approach that uses univariate stable distributions for each asset and a copula for cross-section dependency modeling, combined with expected tail loss (ETL), was initially described in Martin et al. (2003) and subsequently published in extended form in Rachev et al. (2007). See also Rachev et al. (2009) in this volume. ETL is the average loss below value-at-risk (VaR) and as such is a more informative risk than VaR. ETL, which is also known as conditional value-at-risk (CVaR), has the very attractive feature that the mean-ETL optimization problem is convex and can be easily solved with linear programming software, even for large portfolios (Rockafellar and Uryasev 2000).

Mean-ETL portfolio optimization and the risk management approach based on stable distributions, skewed t-distributions, skewed t-copula, and other sophisticated financial econometric modeling methods have been fully developed by Rachev and colleagues and the research team at FinAnalytica, Inc., and implemented in the Cognity risk management and portfolio optimization product. Evidence to date clearly indicates that the approach can lead to superior risk-adjusted portfolio performance and risk management practice in the presence of the extreme market and asset returns behavior that has become common place. We remark that it is relatively easy to find examples where the MVO symmetric variance risk measure penalizes portfolio performance relative to mean-ETL optimization during periods of upside volatility. Since ETL with typical tail probabilities of 0.01 or 0.05 is a more sensitive measure of extreme tail risk than volatility, it is also intuitive that MVO can be slower than mean-ETL optimization in deallocating from portfolio assets that suddenly develop considerable downside risk.

Another way of handling this problem has been developed by researchers A. Clark, M. Labovitz, and H. Turowski at Thomson Reuters, who were in search of ways of generating "alpha" based on the holdings of stock mutual funds. Using results from Malevergne and Sornette (2004a, b), the researchers fit Weibull distributions to the return series and then, using a transform due to Feynman, generated near-normal distributions with good moment generating properties. Standard MVO and ETL tools could then be used at that point (Clark and Labovitz 2006). Alternative approaches would be to model the returns using fat-tailed skewed distributions as above or *long-tailed* distributions, i.e., distributions with tails heavier than the normal but thinner than polynomial tails. (The exponential and stretched exponential are examples of long-tailed distributions.) These researchers found the latter to be more useful. The Thomson Reuters researchers also used the same Rockafellar and Uryasev (2000) ETL method after fitting the tails to optimize the fund holdings.

For all of these techniques, there remains the issue of having "sufficient assets returns histories" for reliable fitting of fat-tailed distributions. For portfolios based on at least a year of daily data, it appears that reliable fits of such distributions is assured [see for example the QQ-plots of daily small-cap returns in Martin et al. (2003)]. This may even be possible with, for instance, 5 years of monthly data for some asset classes (where, however, three-component normal mixture distributions will often suffice). In any event, bootstrap resampling should be used to assess the variability of the estimated parameters in fitting fat-tailed skewed distributions.

On the other hand, normal QQ-plots of monthly returns of portfolio assets will sometimes reveal only one or a small number of outliers and will clearly have insufficient data to reliably fit a fat-tailed skewed distribution. This is the case, for example, with the 60 months of returns for all four of the small-caps in Fig. 11.1. In the case of the hedge fund returns in Fig. 11.2, it appears that there may be enough data in the tails of the returns of F2, F4, and the right tail of F5 to reliably fit a fat-tailed distribution. Similar comments apply to the commodities portfolio monthly returns of Fig. 11.24 and to lesser extent to the ETFs portfolio daily returns of Fig. 11.25. In such cases, robust MVO methods can serve quite a useful purpose, providing at minimum diagnostic guidance (as discussed in Sect. 11.7) and in some cases resulting in a portfolio with better performance than MVO.

Even when reliable fitting of fat-tailed marginal distributions is achievable, there is a complementary use of robust methods in fitting covariance matrices for linear cross-sectional dependency structures and in fitting t-copula models. In the latter case, it is necessary to estimate the parameters of a multivariate t-distribution, and

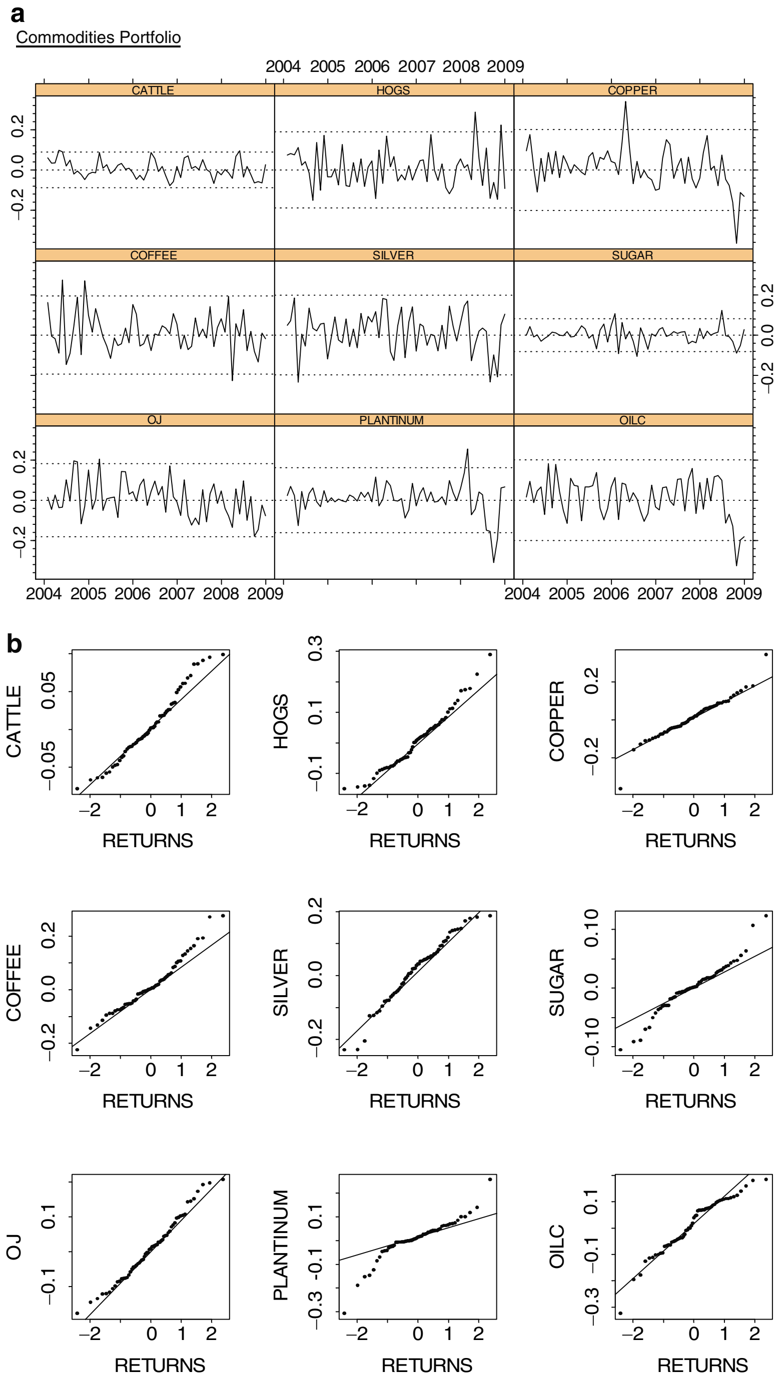


Fig. 11.24 (**a**) Time series of commodities returns. (**b**) QQ-plots of commodities returns. (**c**) Scatter plots of commodities returns

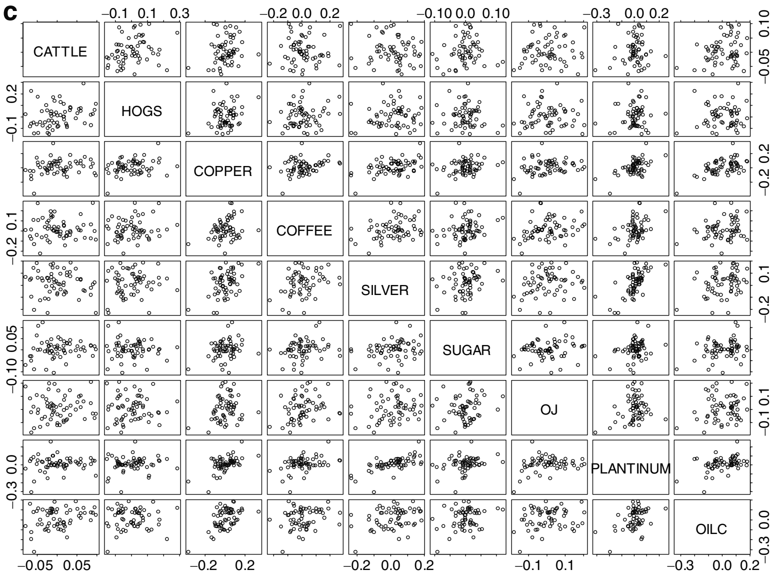


Fig. 11.24 (continued)

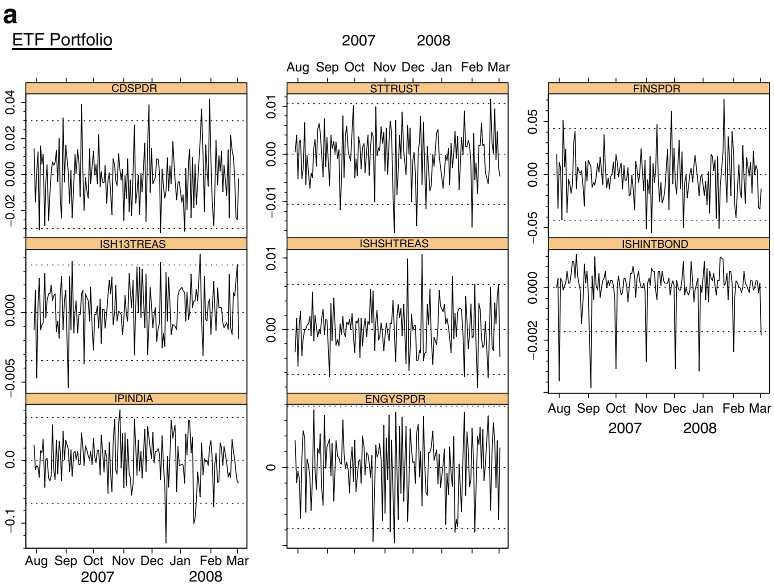
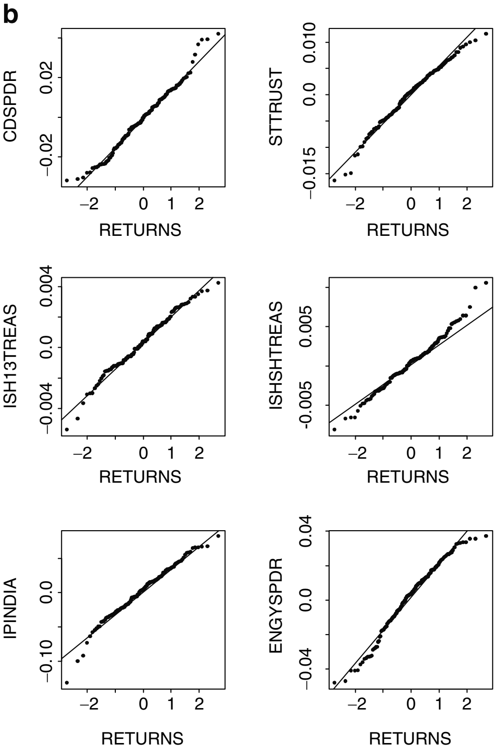


Fig. 11.25 (**a**) Time series of ETF returns. (**b**) QQ-plots of ETF returns. (**c**) Scatter plots of ETF returns

b



c

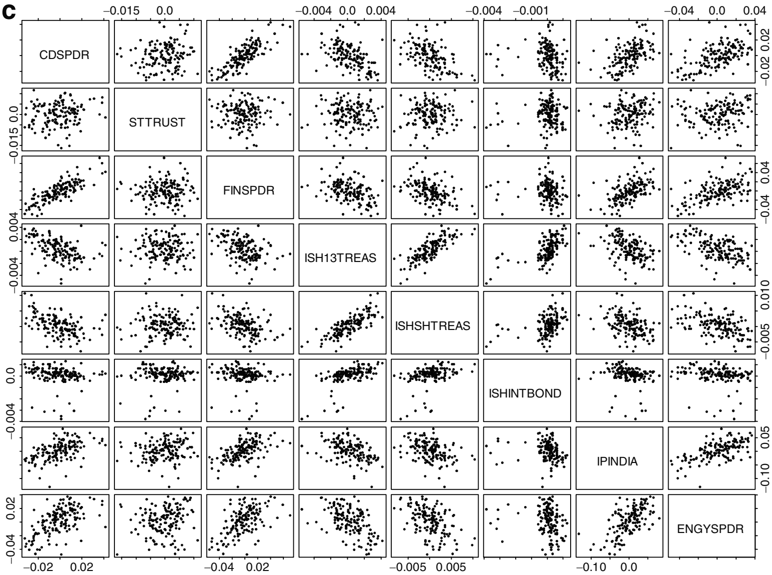


Fig. 11.25 (continued)

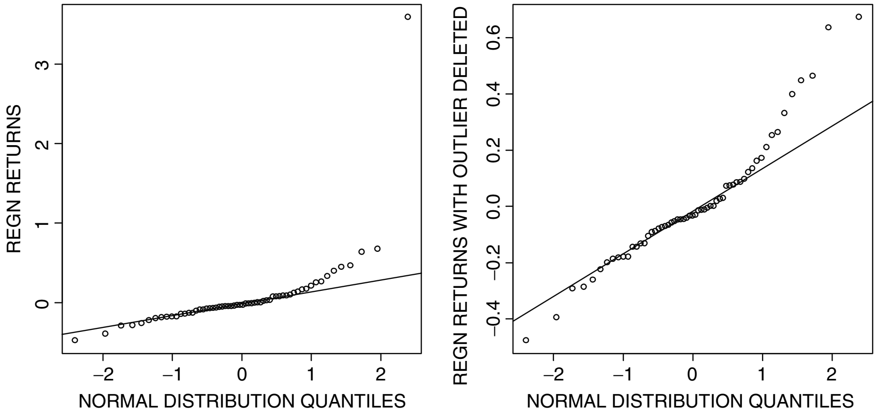


Fig. 11.26 REGN right fat tail with and without wild outlier

the time honored maximum-likelihood estimates (MLEs) of such parameters comprise a class of robust estimates of multivariate scatter (Maronna 1976; Maronna et al. 2006).

Finally, it is possible for a truly wild outlier to bias an estimate (MLE or otherwise) of the parameters of fat-tailed skewed distribution toward a more fat-tailed or skewed distribution than that of the vast majority of the data. Figure 11.26 for the REGN small-cap returns vividly demonstrates the issue (and in this case one might reliably estimate the fat right tail with the outlier deleted, e.g., with a two-component normal mixture model). This suggests that there is a natural role for robust estimation of fat-tailed distributions. So-called "black swan" events are not limited to small-cap stocks, sector- or market-wide events. They have been seen repeatedly in log rank vs. log amplitude plots for bonds, commodities, hedge funds, and stocks other than small caps. Current thinking among some, if not many, quantitative modelers is that these "black swans" are from a separate distribution entirely, one with insufficient data to model effectively.

One should want to get an estimate of the fat-tailed distributions of the assets in their portfolio without the very rare wild outlier ("black swan"), in order to manage the portfolio most of the time, as well as a returns distribution that accounts for the "black swan." The idea that robust estimation may be compatible with extreme value distribution estimation was recently introduced by Dell'Aquila and Embrechts (2006).

11.9 Other Robust Methods in Portfolio Construction

There are a number of other applications of robust statistical methods in portfolio construction; we mention just a few here.

11.9.1 Robust Modified VaR for Portfolio Optimization

As mentioned in Sect. 11.1, one approach to portfolio optimization that adjusts for the nonnormality of the returns distributions is called "modified value-at-risk (MVaR)" based on a Cornish–Fisher four-moment expansion. Unfortunately this approach suffers from two problems. The first is that the Cornish–Fisher expansion can fail to be accurate when far in the tails, e.g., when a commonly used 1% tail probability VaR is used and/or the distribution is far from normal. The second problem is a widely ignored one in the finance literature on uses of skewness and kurtosis, namely that the sample third and fourth moment estimates have variances that explode rapidly under even modest fat-tailed deviations from normality. See, however, Kim and White (2004), who noted the lack of robustness of the classical estimates, studied the use of robust estimates of skewness and kurtosis and urged caution in using the classical sample estimates. Zhu and Martin (2008) studied various robust skewness and kurtosis estimates further, finding that simple trimming estimates behave well. They then used the simple trimming estimates in the skewness and kurtosis terms of the Cornish–Fisher expansion to obtain a robust modified VaR (RMVaR) risk measure for use in portfolio optimization. Their results indicate that RMVaR results on average improved portfolio performance relative to use of the standard MVaR risk measure.

11.9.2 Robust Fundamental Factor Models

There is clearly a role for robust regression in fitting factor models. For example, Knez and Ready (1997) showed that in a Fama-French factor model a very small fraction of outliers can severely bias the least-squares coefficient estimates, resulting in the conclusion that equity returns are negatively related to firm size. They showed that use of a least-trimmed squares robust regression resulted in the opposite conclusion that equity returns are positively related to firm size for the vast majority of the firms, and a small fraction of small-sized firms have outlier returns. This is a more accurate and sensible description of the relationship, and it was discovered by robust regression, not by least squares regression. Motivated by the Knez and Ready result, Martin and Simin (2003) demonstrated the efficacy of robust regression for estimating beta in the single factor model. It remains to carry out an in-depth study of the use of robust regression in fitting fundamental factor models for equity portfolio management, and the results of 6.2 represent a preliminary step toward this goal.

An important question to be answered is: "How frequent and how severe are outliers in the risk factors, and will one-dimensional outlier treatment suffice?" The results of Sect.11.6.2 reveal that even after one-dimensional Winsorizing each individual set of exposures, there remain significant outliers. This means that when fitting a fundamental factor model, a robust fitting method is needed that can cope with outliers in the exposures (the "independent" variables in the regression) as well as in the asset returns (the "dependent" variable in the regression). In the statistical robustness literature, outliers in the independent variables are called "leverage" points. It is known that when the data contains leverage points it does not suffice to use the robust regression M-estimate with an unbounded loss function to obtain bias robustness and that instead a bounded loss function should be used (see Martin et al. 1989 and Maronna et al. 2006). Indeed the optimal bias robust estimate of Svarc et al. (2002) is likely to be effective for fitting equity fundamental factor models.

11.9.3 Robust EWMA and GARCH Models

EWMA or GARCH models are typically needed to deal with pervasive volatility clustering behavior of asset returns. It is a widely overlooked fact that, as a consequence of implementing a filtering operation on returns, EWMA and GARCH models often overestimate volatility following the occurrence of an isolated returns outlier (as opposed to an approximate level shift in volatility). A striking example of this behavior in the EWMA case, along with a simple robust filtering remedy, is provided in Section 6.4.2 of Scherer and Martin (2005). By now a number of research results are available on robust ARCH and GARCH models. See Muler and Yohai (2007) for recent results on robust GARCH models and the references therein to earlier work on robust ARCH and GARCH models. Robust GARCH methods deserve to be extended to the case of multivariate returns and should be used more widely.

11.10 Concluding Comments and Research Directions

We have introduced robust estimates of means, covariance, and correlations for the purpose of portfolio returns correlations and covariance analysis and for robust MVO portfolio applications. Application of the methods to four portfolios representing four diverse asset classes has demonstrated the efficacy of the methods and the pervasive existence of influential outliers across asset classes. The primary focus has been on the diagnostic value delivered by differences between the robust and classical methods for assessing (a) pairwise portfolio asset diversification opportunities as indicated by correlations; (b) overall covariance structure as indicated by eigenvalue structure; and (c) MVO efficient frontier risk and return characteristics. We note that the robust covariance and correlation methods, robust MVO portfolio construction, and associated visualization displays used throughout this paper are implemented in the Cognity portfolio optimization and risk management system (see http://www.finanalytica.com).

In this paper, we have concentrated on small portfolios, i.e., small numbers of assets in the portfolio, as a first step in studying the problem of outlier influence and the use of robust MVO. It remains to be determined if the results presented here carry over to larger portfolios, including equity portfolios selected from universes of up to a few thousand assets. Here two problems need to be addressed. The first is the scalability of robust covariance estimates for large dimensions and the use of deterministic as well as stochastic computation algorithms in cases where a dimension reducing factor model is not used. The second problem is the question of efficacy of using bias robust factor models for dimension reduction, namely, the problems of scalability of robust fitting of factor models for large fundamental factor models (in the case of equity models) and large time series factor models (in the case of fund-of-funds portfolios) and the use of deterministic as well as stochastic computation algorithms to build these models.

Whether robust MVO is to be preferred to classical MVO portfolio construction as a matter of routine practice is a question that remains to be answered by extensive out-of-sample back-testing and application in practice, using weight constraints that are realistic in practice (the latter need to be used in extending the efficient frontier studies of the current paper). A detailed comparison with mean-ETL optimal portfolios based on fat-tailed skewed distribution models is also needed. Studies by the FinAnalytica, Inc. research team suggest that mean-ETL portfolios will deliver superior performance to classical MVO portfolios in situations where one has a substantial amount of weekly or daily returns available. It is to be noted that Clark and Labovitz (2006) show the superior performance of higher moment models against MVO and Fama-French three and four factor models.

References

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. and Tukey, J. W. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton University Press, Princeton, NJ.

Ceria, S. and Stubbs, R. A. (2006). “Incorporating estimation errors into portfolio selection: Robust portfolio construction”, *Journal of Asset Management*, 8, 109–127.

Chernobai, A. and Rachev S. (2006). “Applying robust methods to operational risk modeling”, *Journal of Operational Risk*, 1(1), 27–42.

Chopra, V. and Ziemba, W. (1993). “The effect of means, variances and covariances on optimal portfolio choice”, *Journal of Portfolio Management*, 19(2), 6–11.

Clark, A and Labovitz, M. (2006). “Securities selection and portfolio optimization: Is money being left on the table?”, in *Financial Engineering and Applications – FEA 2006*, Holder M. (ed). Acta Press, Canada.

Cornish, E. and Fisher, R. (1960), “The percentile points of distributions having known cumulants”, *Technometrics*, 2(2), 209–225.

Croux, C. and Haesbroeck, G. (1999). “Influence function and efficiency of minimum covariance determinant scatter matrix estimator”, *Journal of Multivariate Analysis*, 71, 161–190.

Dell’Aquila, A. and Embrechts, P. (2006). “Extremes and robustness: A contradiction?”, *Financial Markets and Portfolio Management*, 20, 103–118.

DeMiguel, V. and Nogales, F. J. (2006). *Portfolio Selection With Robust Estimates of Risk*. Working Paper, Decision Sciences, London Business School.

Fabozzi, F. J., Kolm, P. N., Pachamanova, D. and Focardi, S. M. (2007). *Robust Portfolio Optimization and Management*, Frank J. Fabozzi Series. Wiley, New York.

Favre, L. and Galeano, J. (2002). “Modified value-at-risk optimization with hedge funds”, *Journal of Alternative Investments*, 5(2), 16–28.

Gao, W. (2004). *Portfolio Optimization Based on Robust Estimation Procedures*. M.S. Thesis for Professional Masters Degree in Financial Mathematics, Worcester Polytechnic Institute.

Hampel, F. R., Roncheteti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

Harvey, C., Liechty, J., Liechty, M. and Mueller, P. (2004). *Portfolio Selection With Higher Moments*. Working paper, Department of Marketing and Statistics, Pennsylvania State University.

Huber, P. J. (1964). "Robust estimation of a location parameter", *The Annals of Mathematical Statistics*, 34, 73–101.

Hampel, F. R. (1974). "The influence curve and its role in robust estimation", *The Annals of Statistics*, 69, 383–393.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

Jurczenko, E. and Maillet, B. (2006). "Introduction to multi-moment asset allocation and pricing models", in *Multi-Moment Asset Allocation and Pricing Models*, Jurczenko-Maillet (ed), 21–24. Wiley, New York.

Knez, P. J. and Ready, M. J. (1997). "On the robustness of size and book-to-market in cross-sectional regressions", *Journal of Finance*, 52, 1355–1382.

Kim, T. and White, H. (2004). "On more robust estimates of skewness and kurtosis", *Finance Research Letters*, 1, 56–73.

Kraus, A. and Litzenberger, R. (1976). "Skewness preference and the valuation of risky assets", *Journal of Finance*, 31, 1085–1100.

Lauprete, G. J., Samarov, A. M. and Welsch, R. E. (2002). "Robust portfolio optimization", *Metrika*, 55, 139–149.

Martin, R. D. and Simin, T. (2003). "Outlier-resistant estimates of Beta", *Financial Analysts Journal*, 59(5), 59–69.

Malevergne, Y. and Sornette, D. (2004a). "Multivariate Weibull distributions for asset returns: I", *Finance Letters*, 2(6), 16–32.

Malevergne, Y. and Sornette, D. (2004b). "High-order moments and cumulants of multivariate Weibull asset returns distributions: analytical theory and empirical tests: II", *Finance Letters*, 3(1), 54–63.

Maronna, R. A. (1976). "Robust M-estimators of multivariate location and scatter", *The Annals of Statistics*, 4, 51–67.

Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley, New York.

Martin, R. D., Rachev, S. and Siboulet, F. (2003). "Phi-alpha optimal portfolios and extreme risk management", *Wilmott Magazine*, November/December Issue.

Martin, R. D., Yohai, V. J. and Zamar, R. H. (1989). "Min–max bias robust regression", *The Annals of Statistics*, 17, 1608–1630.

Martin, R. D. and Zamar, R. H. (1993a). "Bias robust estimates of scale", *The Annals of Statistics*, 21, 991–1017.

Martin, R. D. and Zamar, R. H. (1993b). "Efficiency-constrained bias-robust estimation of location", *The Annals of Statistics*, 21, 338–354.

Muler, N. and Yohai, V. J. (2007). "Robust estimates for GARCH models", *Journal of Statistical Planning and Inference*, 138(10), 2918–2940.

Perret-Gentil, C. and Victoria-Freser, M. P. (2004). *Robust Mean–Variance Portfolio Selection*. Working Paper No. 173, National Center of Competence in Research, Financial Valuation and Risk Management, Swiss National Science Foundation.

Rachev, S., Martin, R. D., Racheva, B. and Stoyanov, S. (2007). "Stable ETL optimal portfolios and extreme risk management", in *Risk Assessment: Decisions in Banking and Finance*, Bol, Rachev and Wuerth (eds). Springer, New York.

Rachev, S., Menn, C. and Fabozzi, F. (2005). *Fat-Tailed and Skewed Asset Return Distributions: Implications for Risk Management, Portfolio Selection and Option Pricing*. Wiley, New York.

Rachev, S. and Mittnik, S. (2000). *Stable-Paretian Models in Finance*. Wiley, New York.

Rachev, S., Rachev-Iotova, B., Stoyanov, S. and Fabozzi, F. (2009). "Risk management and portfolio optimization for volatile markets", in *The Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, Guerard J. B. (ed). Springer, New York.

Rockafellar, R. T. and Uryasev, S. (2000). "Optimization of Conditional Value-at-Risk", *Journal of Risk*, 3, 21–41.

Rousseeuw, P. J. (1985). "Multivariate estimation with high breakdown point", in *Mathematical Statistics and its Applications (vol. B)*, Grossmann W., Pflug G., Vincze, I. and Wertz, W. (eds.), 283–297. Reidel, Dordrecht.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

Rousseeuw, P. J. and van Driessen, K. (1999). "A fast algorithm for the minimum covariance determinant estimator", *Technometrics*, 41, 212–223.

Scherer, B. (2007). "Can robust portfolio optimization help to build better portfolios?" *Journal of Asset Management*, 7(6), 374–387.

Scherer, B. (2007). *Portfolio Construction and Risk Budgeting*, 3rd extended edition. Risk Books, London.

Scherer, B. and Martin, R. D. (2005). *Introduction to Modern Portfolio Optimization with NUOPT and S-PLUS*. Springer, New York.

S-PLUS 6 Robust Library User Guide, Version 1.0 (2002). Tibco, Inc.

Svarc, M., Yohai, V. J. and Zamar, R. H. (2002). "Optimal bias-robust M-estimates of regression", in *Statistical Data Analysis Based on the L1 Norm and Related Methods*, Dodge Y (ed), 191–200. Birkhäuser, Basle.

Tukey, J. W. (1960). "A survey of sampling from contaminated distributions", in *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, Olkin I. et al. (eds), 448–485. Stanford University Press, Palo Alto, CA.

Welsch, R. E. and Zhou, X. (2007). "Applications of robust statistics to asset allocation models", *REVSTAT – Statistical Journal*, 5(1), 97–114.

Yohai, V. J. (1987). "High breakdown-point and high efficiency estimates for regression", *The Annals of Statistics*, 15, 642–656.

Yohai, V. J. and Zamar, R. H. (1997). "Optimal locally robust M-estimators of regression", *Journal of Statistical Planning and Inference*, 64, 309–323.

Zhu, M. and Martin, R. D. (2008). *Robust Modified VaR Portfolio Optimization*. Preprint, Department of Statistics, University of Washington.

Part II
Owitz and the Expanding Definition of Risk: Applications of Multi-Factor Risk Models

Chapter 12
Applying Markowitz's Critical Line Algorithm

Andras Niedermayer and Daniel Niedermayer

12.1 Introduction

In recent years, Monte Carlo methods used for portfolio resampling have become increasingly popular in financial engineering and research.[1] When lower and upper bounds on portfolio holdings are imposed, these methods repeatedly employ quadratic optimization algorithms and, therefore, require a fast implementation of the mean variance optimizer. However, fast implementations are not publicly available. This is surprising given that with some simple numerical improvements an implementation of Markowitz's (1952) original Critical Line Algorithm (CLA) significantly outperforms standard software packages and a recently developed quadratic optimization algorithm described in Steuer et al. (2006). Therefore, the aim of this article is to derive all steps needed by the CLA and also to provide a didactic alternative to the framework in Markowitz and Todd (2000). The derivation of the equations and the pseudocode in the Appendix should ease the implementation of the fast version of the CLA by researchers and financial practitioners alike.

As a benchmark for computational speed, we use the results in Steuer et al. (2006). They develop a simplex-based algorithm that calculates all turning points of the constrained minimum variance frontier (CMVF) while significantly reducing computational time compared to standard software packages such as Matlab, Cplex, LINGO, Mathematica, and premium Solver. In order to compare their results with the CLA, we implement a numerically enhanced version of the algorithm in Fortran 90. We show that this algorithm outperforms the algorithm in Steuer et al. (2006) by a factor of almost 10,000 (for 2,000 assets) and standard software packages by even more.

A. Niedermayer (✉)
Kellogg School of Management, Managerial Economics and Decision Sciences, Northwestern University, Evanston, IL 60208, USA
e-mail: a-niedermayer@kellogg.northwestern.edu

D. Niedermayer
Credit Suisse, Asset Management and WWZ, University of Basel, Holbeinstrasse 12, CH-4051 Basel, Switzerland

[1] Resampling simulations have been introduced in, e.g., Jorion (1992) and Michaud (1998) and studied in, e.g., Scherer (2002), Markowitz and Usmen (2003), Scherer and Martin (2006), and Wolf (2006).

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_12,
© Springer Science+Business Media, LLC 2010

From this observation, we conclude that the high performance of the CLA is not well known. In fact, excluding the paper by Steuer et al. (2006), no studies benchmarking quadratic optimization algorithms' performance are known to us. Moreover, as no publicly available software package exists that computes the entire CMVF, we provide a Matlab optimization package using our Fortran 90 implementation of the CLA.[2]

Finally, this paper can be considered as a didactic alternative to the standard CLA as presented by Markowitz. All numerical improvements to the algorithm are treated explicitly.

The rest of the paper is organized as follows: Sect. 12.2 introduces the mathematical framework and definitions required by the CLA. Section 12.3 formulates the quadratic optimization method and the numerical improvement. Section 12.4 describes performance tests and computational experience. Section 12.5 presents the conclusions. Proofs and a pseudocode representation are given in the Appendix.

12.2 The Framework

Given is a universe of n assets whose returns have the expected value $\boldsymbol{\mu}$ (n vector) and the $n \times n$ positive definite covariance matrix $\boldsymbol{\Sigma}$. In the following, we will denote an investor's weights assigned to each asset by the n vector $\boldsymbol{w}$ with the components summing up to 1.

For a minimum variance portfolio where lower and upper bounds on asset weights are included, we define an n vector containing the asset weights' lower bounds ($w_i \geq l_i, \quad \forall i$) $\boldsymbol{l}$; an n vector containing the asset weights' upper bounds ($w_i \leq u_i, \quad \forall i$) $\boldsymbol{u}$; and $\mathbb{F}$, a subset of $N = \{1, 2, \ldots, n\}$ containing all assets' indices where weights are within their bounds ($l_i < w_i < u_i$) and its size $k \equiv |\mathbb{F}|$. We shall call the corresponding assets as *free assets*. We further define $\mathbb{B}$, the subset of all asset indices where the weights lie on one of their bounds ($w_i = l_i$ or $w_i = u_i$). Thus, $\mathbb{B} = \{1, 2, \ldots, n\} \setminus \mathbb{F}$. The sets of assets on their upper and lower bounds will be called $\mathbb{U}$ and $\mathbb{L}$, respectively, with $\mathbb{B} = \mathbb{U} \cup \mathbb{L}$.

When writing the free assets' indices at the beginning, the covariance matrix $\boldsymbol{\Sigma}$, the expected return vector $\boldsymbol{\mu}$, and the weight vector $\boldsymbol{w}$ can be subdivided into

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma_F} & \boldsymbol{\Sigma_{FB}} \\ \boldsymbol{\Sigma_{BF}} & \boldsymbol{\Sigma_B} \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu_F} \\ \boldsymbol{\mu_B} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{w} = \begin{bmatrix} \boldsymbol{w_F} \\ \boldsymbol{w_B} \end{bmatrix} \tag{12.1}$$

with the covariance matrices $\boldsymbol{\Sigma_F}$ and $\boldsymbol{\Sigma_B}$ of dimensions $(k \times k)$ and $(n-k) \times (n-k)$, the matrices $\boldsymbol{\Sigma_{FB}}$ and $\boldsymbol{\Sigma_{BF}}$ of dimensions $k \times (n-k)$ and $(n-k) \times k$, two k vectors $\boldsymbol{\mu_F}$ and $\boldsymbol{w_F}$, and two $(n-k)$ vectors $\boldsymbol{\mu_B}$ and $\boldsymbol{w_B}$. Moreover, from the symmetry of $\boldsymbol{\Sigma}$ follows $\boldsymbol{\Sigma_{BF}} = \boldsymbol{\Sigma}'_{\boldsymbol{FB}}$.

[2] We also provide an interface for the open source Numerical Python system on Linux.

12.2.1 Unconstrained Case

Before turning to the constrained variance minimization, it is worth to familiarize oneself again with the unconstrained portfolio minimization problem. The unconstrained problem can be solved by means of the Lagrange function

$$L = \frac{1}{2}\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} - \gamma(\boldsymbol{w}'\mathbf{1} - 1) - \lambda(\boldsymbol{w}'\boldsymbol{\mu} - \mu_{\mathrm{p}}), \tag{12.2}$$

with the Lagrange coefficients γ and λ and the target expected return level μ_{p}.[3] The first constraint in (12.2) ensures that assets' weights sum to one; the second constraint tells that the portfolio's expected return equals the target level μ_{p}. Differentiating with respect to $\boldsymbol{w}$, γ, and λ and setting the results to zero, one obtains a system of $(n+2)$ linear equations. Solving this system leads to the solution of the variance minimizing weight vector $\boldsymbol{w}^*$. Obviously, $\boldsymbol{w}^*$ will not generally satisfy the constraints $l_i \leq w_i \leq u_i$.

12.2.2 Constrained Case

The computation of efficient portfolios becomes more difficult when inequality constraints on asset holdings are included. If short selling of assets is forbidden, asset weights must be nonnegative and the constraint has the form of $\boldsymbol{w} \geq 0$.

However, some problems require a more general constraint with upper and lower bounds. For such problems, the optimal solution will be a portfolio where assets' weights lie within their respective bounds, thus, $l_i \leq w_i \leq u_i$ for all i. In the following, we shall call the solution of this problem a constrained minimum variance portfolio[4] and the graphical representation of the set of solutions in the $(\mu_{\mathrm{p}}, \sigma_{\mathrm{p}})$ plane as CMVF.

One important feature of the CMVF is the existence of turning points.[5]

Definition 12.1. A constrained minimum variance portfolio is called *turning point* if other constrained minimum variance portfolios in its vicinity contain different free assets. □

When knowing which assets at a certain expected return level μ_{p} are free in the constrained minimum variance portfolio, thus, knowing $\mathbb{F}$, the problem can be formulated easily. This is stated in the following proposition.

---

[3] In the following, we will denote a vector of ones $(1, \ldots, 1)'$ as $\mathbf{1}$. To emphasize that the size of a vector $\mathbf{1}$ is the same as of a set $\mathbb{A}$, we will write $\mathbf{1}_{\boldsymbol{A}}$ (e.g., $\mathbf{1}_{\boldsymbol{F}}$ has size k and $\mathbf{1}_{\boldsymbol{B}}$ has size $n-k$).

[4] This is also called feasible mean variance efficient portfolio in the literature.

[5] In Markowitz (1959), turning points are described as the intersections of two critical lines.

Proposition 12.1. *The free weights in the solution* $\boldsymbol{w}^*$ *of the constrained case are equal to the weights in the solution of the unconstrained case with the Lagrange function*

$$\begin{aligned} L =& \frac{1}{2}\begin{bmatrix} \boldsymbol{w_F} \\ \boldsymbol{w_B} \end{bmatrix}' \begin{bmatrix} \boldsymbol{\Sigma_F} & \boldsymbol{\Sigma_{FB}} \\ \boldsymbol{\Sigma_{BF}} & \boldsymbol{\Sigma_B} \end{bmatrix} \begin{bmatrix} \boldsymbol{w_F} \\ \boldsymbol{w_B} \end{bmatrix} \\ & - \gamma \left(\begin{bmatrix} \boldsymbol{w_F} \\ \boldsymbol{w_B} \end{bmatrix}' \begin{bmatrix} \boldsymbol{1_F} \\ \boldsymbol{1_B} \end{bmatrix} - 1 \right) - \lambda \left(\begin{bmatrix} \boldsymbol{w_F} \\ \boldsymbol{w_B} \end{bmatrix}' \begin{bmatrix} \boldsymbol{\mu_F} \\ \boldsymbol{\mu_B} \end{bmatrix} - \mu_{\mathrm{p}} \right) , \end{aligned} \tag{12.3}$$

where the components $\boldsymbol{w_F}$ *are subject to minimization and the components* $\boldsymbol{w_B}$ *are fixed to their actual values.*

Proof. *This is obvious: changing* $\boldsymbol{w_F}$ *infinitesimally ensures that all weights remain free. This cannot lead to a smaller* L *otherwise* $\boldsymbol{w}^*$ *would not be a solution of the constrained case.* □

There is a major difference between (12.2) and (12.3); the underlying subsets in (12.3) depend on the constrained minimum variance portfolio's location. Since the subsets $\mathbb{F}$ and $\mathbb{B}$ do not change between turning points, the solutions between two turning points will be the solution of an unconstrained optimization upon the subset $\mathbb{F}$.

Corollary 12.1. *Combining two neighboring turning points with a real weight* $\omega \in [0, 1]$ *always leads to a constrained minimum variance portfolio.*

Proof. *This follows from Proposition 12.1 and the fact that a linear combination of two solutions (for different values of* μ_{p}*) of the unconstrained problem is a solution as well.* □

Differentiating (12.3) with respect to $\boldsymbol{w_F}$ yields

$$\boldsymbol{\Sigma_F w_F} + \boldsymbol{\Sigma_{FB} w_B} - \lambda \boldsymbol{\mu_F} = \gamma \boldsymbol{1_F} \, . \tag{12.4}$$

At this point the constraint $\boldsymbol{w'1} = 1$ must be adapted according to (12.1) leading to

$$\boldsymbol{1'_F w_F} = 1 - \boldsymbol{1'_B w_B} \, . \tag{12.5}$$

Solving (12.4) together with (12.5) for γ yields

$$\gamma = -\lambda \frac{\boldsymbol{1'_F \Sigma_F^{-1} \mu_F}}{\boldsymbol{1'_F \Sigma_F^{-1} 1_F}} + \frac{1 - \boldsymbol{1'_B w_B} + \boldsymbol{1'_F \Sigma_F^{-1} \Sigma_{FB} w_B}}{\boldsymbol{1'_F \Sigma_F^{-1} 1_F}} \, . \tag{12.6}$$

Note that here λ is set exogenously instead of μ_{p}. Therefore, λ determines the value of γ and finally the expected return of the minimum variance portfolio. The value of μ_{p} is fictitious in (12.3) and the optimal solution is solely determined by λ. In fact, it is very similar to set λ exogenously or to calculate with a fixed μ_{p} and look at λ as Lagrange multiplier. This is because λ and $\boldsymbol{\mu' w}$ $(= \mu_{\mathrm{p}})$ are linearly related between turning points and because a higher λ yields a (constrained) minimum variance portfolio with higher expected return. This is stated here by Proposition 12.2 with the proof being banned to the Appendix.

Proposition 12.2. *Between two turning points, λ and $\boldsymbol{\mu}'\boldsymbol{w}$ are linearly related with a positive slope*

$$\frac{\partial(\boldsymbol{\mu}'\boldsymbol{w}(\lambda))}{\partial\lambda} > 0 . \qquad \square$$

12.3 The Algorithm

The main idea for the algorithm presented is the following: first, the turning point with the highest expected return value is found; then the next lower turning point is calculated. This is illustrated in Fig. 12.1.

From the definition of a turning point, we know that each of them will differ in the composition of its free assets. Therefore, for each of them (12.4) will hold for a different subset $\mathbb{F}$. Except for the case where two or more turning points lie upon each other,[6] when passing a turning point one has to add or remove exactly one element from the set of free assets $\mathbb{F}$.

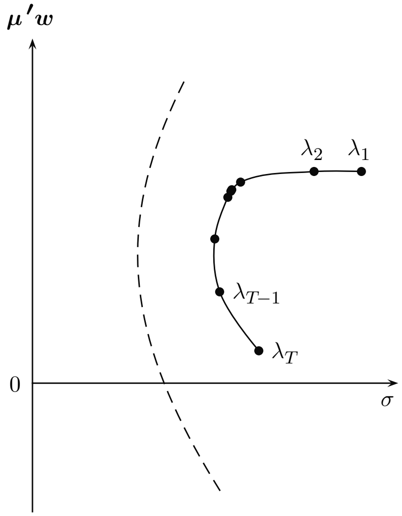


Fig. 12.1 This figure shows the minimum variance frontier (*dashed line*) and the constrained minimum variance frontier for ten assets and an arbitrarily chosen nonsingular covariance matrix. The *dots* represent the constrained frontier's turning points

[6] We do not discuss this possibility even though the algorithm can cope with it; when calculating numerically, there will hardly be two or more turning points on one (σ, μ) location. Markowitz (1959) proposes as a solution to this situation to either alter the μ of one asset slightly or to use the method described in Markowitz (1956).

Moreover, when moving downward from a turning point to the next one, λ will decrease (see Proposition 12.2). When looking at turning points such as in Fig. 12.1 it must, therefore, be that

$$\lambda_1 > \lambda_2 > \lambda_3 > \cdots > \lambda_T$$

with T being the number of turning points.

Sections 12.3.1 and 12.3.2 show how to find the turning point with the highest expected return (turning point 1) and how to move to the next lower turning point. Section 12.3.3 shows a way how to improve the algorithm's performance significantly.

12.3.1 The Starting Solution

This algorithm requires an initial solution on the CMVF. It is convenient to find the turning point with the highest expected return before moving to the next lower turning point.

Therefore, we order all assets with respect to their expected return values such that the first asset has the highest and the last asset the lowest expected return. After setting all asset weights to their lower bounds ($w_i = l_i$), we start to increase the weight of the first asset. If the upper bound is reached and $\boldsymbol{w}'\mathbf{1} < 1$, we start to increase the second assets' weight, and so forth. This procedure terminates when $\boldsymbol{w}'\mathbf{1} = 1$ is reached.[7]

The solution is typically a weight vector where the weights of the first assets are set to their upper bounds and the last assets' weights are set to their lower bounds. There is one asset in the middle, where the weight is between its bounds. We will call this asset the free asset and index it with i_{free}. The weight of the free asset is $1 - \sum_{i \in \mathbb{U}} w_i - \sum_{i \in \mathbb{L}} w_i \equiv 1 - \boldsymbol{w}'_{\boldsymbol{B}} \mathbf{1}_{\boldsymbol{B}}$.

We solve a simpler problem than Markowitz and Todd (2000) as we have the only equality constraint $\boldsymbol{w}'\mathbf{1} = 1$, whereas Markowitz and Todd consider the general constraint $\boldsymbol{A}\boldsymbol{w} = \boldsymbol{b}$. However, because in many practical situations the specific case $\boldsymbol{w}'\mathbf{1} = 1$ is sufficient, one can use our simpler algorithm to find the first turning point instead of the simplex algorithm used in Markowitz and Todd (2000).

12.3.2 Iteration

When moving from a turning point to the next lower one by decreasing λ, one of the following two situations will occur; either one free asset moves to one of its

[7] Obviously, this requires $\mathbf{1}'\boldsymbol{l} \leq 1 \leqslant \mathbf{1}'\boldsymbol{u}$. For $\mathbf{1}'\boldsymbol{u} < 1$ or $\mathbf{1}'\boldsymbol{l} > 1$, there is no solution to the portfolio optimization problem. For $\mathbf{1}'\boldsymbol{u} = 1$ or $\mathbf{1}'\boldsymbol{l} = 1$, the whole efficient frontier consists of only one portfolio, namely $\boldsymbol{w} = \boldsymbol{u}$ or $\boldsymbol{w} = \boldsymbol{l}$, respectively.

bounds or an asset formerly on its bound becomes free. These two cases have to be considered in order to compute the next turning point's λ and $\boldsymbol{w}$.

12.3.2.1 Case (a) One Formerly Free Asset Moves to Its Bound

Let λ_{current} belong to a turning point and let $\mathbb{F}$ be the set of the free assets slightly below this turning point (i.e., for λ such that $\lambda_{\text{current}} \equiv \lambda_t > \lambda > \lambda_{t+1}$).

For this subset containing k variables (12.4) holds and thus

$$\boldsymbol{w_F} = -\boldsymbol{\Sigma_F^{-1}\Sigma_{FB}w_B} + \gamma\boldsymbol{\Sigma_F^{-1}1_F} + \lambda\boldsymbol{\Sigma_F^{-1}\mu_F}\,. \tag{12.7}$$

Substituting γ from (12.6) into (12.7) gives us $\boldsymbol{w_F}$ as a linear function of λ. When decreasing λ, the asset i will hit the lower bound if the derivative $\mathrm{d}w_i/\mathrm{d}\lambda$ is positive and hit the upper bound if the derivative is negative. We denote the value of λ as $\lambda^{(i)}$ at the point where asset i hits the corresponding bound. $\lambda^{(i)}$ can be derived from the linear relation between $\boldsymbol{w_F}_i$ and λ resulting from (12.6) and (12.7). This gives

$$\begin{aligned}\lambda^{(i)} =& \frac{1}{C_i}\Big[\big(1-\boldsymbol{1'_B w_B}+\boldsymbol{1'_F\Sigma_F^{-1}\Sigma_{FB}w_B}\big)(\boldsymbol{\Sigma_F^{-1}1_F})_i \\ &-(\boldsymbol{1'_F\Sigma_F^{-1}1_F})\big(b_i+(\boldsymbol{\Sigma_F^{-1}\Sigma_{FB}w_B})_i\big)\Big]\end{aligned} \tag{12.8}$$

with

$$C_i = -(\boldsymbol{1'_F\Sigma_F^{-1}1_F})(\boldsymbol{\Sigma_F^{-1}\mu_F})_i + (\boldsymbol{1'_F\Sigma_F^{-1}\mu_F})(\boldsymbol{\Sigma_F^{-1}1_F})_i \tag{12.9}$$

and[8]

$$b_i = \begin{cases} u_i & \text{if } C_i > 0\,, \\ l_i & \text{if } C_i < 0\,. \end{cases} \tag{12.10}$$

Note that for $k = 1$ one gets $C_i = 0$, which reflects the fact that the constraint $\boldsymbol{1'_F w_F} = w_i = 1 - \boldsymbol{1'_B w_B}$ uniquely determines w_i. Therefore, case (a) should be considered only for $k > 1$. C_i is zero for all i if accidentally $\boldsymbol{\mu_F}$ is proportional to $\boldsymbol{1_F}$, i.e., $\mu_i = \mu_j$ for all $i, j \in \mathbb{F}$.

The next $\lambda < \lambda_{\text{current}}$ where an asset wants to leave the subset $\mathbb{F}$ is

$$\lambda_{\text{inside}} = \max_{i\in\mathbb{F}}\{\lambda^{(i)}\}\,, \tag{12.11}$$

or λ_{inside} does not exist if $k = 1$ or $C_i = 0$ for all i.

[8] Note that $\mathrm{d}w_i/\mathrm{d}\lambda = -C_i/(\boldsymbol{1'_F\Sigma_F^{-1}1_F})$. Also note further that $i \in \mathbb{F} = \{i_1, i_2, \ldots, i_k\}$ and, therefore, i can take values from 1 to n (and not from 1 to k).

However, λ_{inside} will only describe the next lower turning point if there is no portfolio with a λ where $\lambda_{\text{current}} > \lambda > \lambda_{\text{inside}}$ and where an asset on its bound wants to get into the subset $\mathbb{F}$. This situation is summarized by case (b).

12.3.2.2 Case (b) One Asset Formerly on Its Bound Wants to Become Free

When moving downward in $\boldsymbol{\mu}'\boldsymbol{w}$ it might occur that an asset i formerly on its bound wants to become free. The corresponding portfolio is, therefore, a turning point where the subsets $\mathbb{F}$ and $\mathbb{B}$ have to be redefined. Let us denote the new subsets as

$$\begin{aligned} \mathbb{F}_i &\equiv \mathbb{F} \cup \{i\}, \\ \mathbb{B}_i &\equiv \mathbb{B} \setminus \{i\}, \end{aligned}$$

where $i \in \mathbb{B}$. Analogously to (12.8) the value $\lambda^{(i)}$ where the newly included asset i's weight moves away from its bound is given by

$$\begin{aligned} \lambda^{(i)} = \frac{1}{C_i} \Big[& \left(1 - \mathbf{1}'_{\boldsymbol{B}_i} \boldsymbol{w}_{\boldsymbol{B}_i} + \mathbf{1}'_{\boldsymbol{F}_i} \boldsymbol{\Sigma}_{\boldsymbol{F}_i}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{F}_i \boldsymbol{B}_i} \boldsymbol{w}_{\boldsymbol{B}_i}\right) (\boldsymbol{\Sigma}_{\boldsymbol{F}_i}^{-1} \mathbf{1}_{\boldsymbol{F}_i})_i \\ & - (\mathbf{1}'_{\boldsymbol{F}_i} \boldsymbol{\Sigma}_{\boldsymbol{F}_i}{}^{-1} \mathbf{1}_{\boldsymbol{F}_i}) \left(b_i + (\boldsymbol{\Sigma}_{\boldsymbol{F}_i}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{F}_i \boldsymbol{B}_i} \boldsymbol{w}_{\boldsymbol{B}_i})_i\right) \Big] \end{aligned} \tag{12.12}$$

with

$$C_i = -(\mathbf{1}'_{\boldsymbol{F}_i} \boldsymbol{\Sigma}_{\boldsymbol{F}_i}^{-1} \mathbf{1}_{\boldsymbol{F}_i})(\boldsymbol{\Sigma}_{\boldsymbol{F}_i}^{-1} \boldsymbol{\mu}_{\boldsymbol{F}_i})_i + (\mathbf{1}'_{\boldsymbol{F}_i} \boldsymbol{\Sigma}_{\boldsymbol{F}_i}^{-1} \boldsymbol{\mu}_{\boldsymbol{F}_i})(\boldsymbol{\Sigma}_{\boldsymbol{F}_i}^{-1} \mathbf{1}_{\boldsymbol{F}_i})_i \tag{12.13}$$

and where for convenience b_i is used for $(\boldsymbol{w}_{\boldsymbol{F}_i})_i = w_i$, which is an asset on its bound possibly becoming free. If asset i was previously on its upper bound, b_i stands for u_i and if it was on the lower bound, it stands for l_i.[9]

In order to find the maximal $\lambda^{(i)} < \lambda_{\text{current}}$ where an asset i currently on its bound wants to become free, (12.12) must be applied for all $i \in \mathbb{B}$.

$$\lambda_{\text{outside}} = \max_{i \in \mathbb{B}} \{\lambda^{(i)} \mid \lambda^{(i)} < \lambda_{\text{current}}\}. \tag{12.14}$$

Again, if no $\lambda^{(i)} < \lambda_{\text{current}}$ exists, we remember that there is no solution for λ_{outside}.

[9] To avoid identifying the trivial solution $\lambda^{(i)} = \lambda_{\text{current}}$ erroneously as a turning point if asset i just went out in the previous step one can check the derivative $\mathrm{d}w_i/\mathrm{d}\lambda$ similarly to case (a). If the derivative is negative and the asset was previously on the upper bound, we have the λ_{current} of the previous turning point and numerical imprecision has led to $\lambda^{(i)} = \lambda_{\text{current}} - \epsilon$ with ϵ small but positive.

12.3.2.3 Finding the Next Turning Point

In order to find out which case will occur, the values of λ_{inside} and λ_{outside} must be compared.

- If solutions for both λ_{inside} and λ_{outside} could be found, then the next turning point will have a λ defined as

$$\lambda_{\text{new}} = \max\{\lambda_{\text{inside}}, \lambda_{\text{outside}}\}\,.$$

  Thus, e.g., case (a) is characterized by $\lambda_{\text{inside}} > \lambda_{\text{outside}}$.
- If a solution only for λ_{inside} or λ_{outside} could be found, λ_{new} is overwritten by the respective value.
- Depending on which case occurs, we replace $\mathbb{F}$ by $\mathbb{F}\setminus\{i\}$ or by $\mathbb{F}_i$, $\mathbb{B}$ by $\mathbb{B}\cup\{i\}$ or $\mathbb{B}_i$, and λ_{current} by λ_{new}.
- If no solution for λ_{inside} and λ_{outside} could be found, we have reached the lowest turning point and the algorithm terminates.[10]

The pseudocode in Appendix summarizes the algorithm for the simplified case when $l_i = 0$ and $u_i = +\infty$.[11]

One way of checking the results is to look at the last turning point's weight vector. This weight vector must be the "opposite" one to the initial solution. When ordering all weights with regard to their expected returns, the last weights must be at their upper and the first weights at their lower bounds. The remaining weight will correspond to the free asset.

Note that in contrast to the calculations in (12.8), the specification of $\mathbb{F}_i$ in (12.12) depends on i and $\boldsymbol{\Sigma}_{\boldsymbol{F}_i}^{-1}$ must be recalculated for each $i \notin \mathbb{F}$.[12] Moreover, as the next turning point has one asset more or one asset less in $\mathbb{F}$, the inverse of the respective covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{F}}^{-1}$ must be recalculated each time. We will show in the following, how these time consuming computations of inverses can be avoided.

12.3.3 Improving Performance

In the algorithm described above, the compositions of $\mathbb{F}$ and $\mathbb{B}$ change with one asset being included or excluded. Here we show that one can avoid recalculating the inverse of the corresponding matrices each time.

---

[10] For practical purposes, the lower half of the efficient frontier (with μ decreasing and σ increasing) does not interest us. In this case, we can terminate the algorithm when σ starts increasing again which corresponds to $\lambda = 0$.

[11] Note that in this case $\boldsymbol{w}_{\boldsymbol{B}} = 0$ and the equations become much simpler. The starting solution is also simpler: one has to set the weight of the asset with the highest expected return to 1.

[12] Or at least the vectors $\boldsymbol{\Sigma}_{\boldsymbol{F}_i}^{-1}\mathbf{1}_{\boldsymbol{F}_i}$ and $\boldsymbol{\Sigma}_{\boldsymbol{F}_i}^{-1}\boldsymbol{\mu}_{\boldsymbol{F}_i}$ have to be calculated.

12.3.3.1 Expansion of the Covariance Matrix $\boldsymbol{\Sigma}_F$

Lemma 12.1. *Let $\boldsymbol{A}$ be a symmetric nonsingular $k \times k$ matrix, $\boldsymbol{a}$ a $k \times 1$ vector, and α be a scalar. Then for the expanded matrix's inverse*

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{a} \\ \boldsymbol{a}' & \alpha \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{A}^{-1} + \beta \boldsymbol{c}\boldsymbol{c}' & -\beta \boldsymbol{c} \\ -\beta \boldsymbol{c}' & \beta \end{bmatrix} \tag{12.15}$$

holds where

$$\boldsymbol{c} = \boldsymbol{A}^{-1}\boldsymbol{a} \qquad \textit{and} \qquad \beta = \frac{1}{\alpha - \boldsymbol{c}'\boldsymbol{a}}.$$

Proof. *Multiplying the expanded matrix with the right-hand side of* (12.15) *yields the identity matrix.* □

Our algorithm requires often expanding the subset $\mathbb{F}$ by one element i and recalculating the inverse of the covariance matrix, $\boldsymbol{\Sigma}_{F_i}^{-1}$, for the new subset. Lemma 12.1 frees us from the burden of making this calculation all over again. This reduces the number of operations for inverting $\begin{bmatrix} \boldsymbol{A} & \boldsymbol{a} \\ \boldsymbol{a}' & \alpha \end{bmatrix}$ from $k^3/3$ to $2k^2$.

12.3.3.2 Reduction of the Covariance Matrix $\boldsymbol{\Sigma}_F$

Reducing the covariance matrix by one row and column does not require the inversion of the newly obtained matrix either. Having calculated the inverse of the expanded covariance matrix as in the previous section and now deleting the given row and column, the newly obtained matrix's inverse can be calculated. This is stated in Lemma 12.2 where for presentational purposes the given index is assumed to be the last one.

Lemma 12.2. *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be $k \times k$ matrices, $\boldsymbol{a}$ and $\boldsymbol{b}$ k vectors, and α and β be two scalars. Then if*

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{a} \\ \boldsymbol{a}' & \alpha \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{B} & \boldsymbol{b} \\ \boldsymbol{b}' & \beta \end{bmatrix} \tag{12.16}$$

holds then

$$\boldsymbol{A}^{-1} = \boldsymbol{B} - \frac{1}{\beta}\boldsymbol{b}\boldsymbol{b}'.$$

holds as well.

Proof. *Combine* (12.15) *and* (12.16) *and solve for $\boldsymbol{A}^{-1}$.* □

12.3.3.3 Key Improvement

The most remarkable improvement stems from the fact that in (12.12) we need to know neither $\boldsymbol{\Sigma}_{F_i}^{-1}$ nor $\boldsymbol{\Sigma}_{F_i}^{-1}\boldsymbol{\Sigma}_{F_i B_i}$. This is stated in Proposition 12.3.

Proposition 12.3. *Expression* (12.12),

$$\lambda^{(i)} = \frac{1}{C_i}\Big[\Big(1 - \mathbf{1}'_{B_i}\boldsymbol{w}_{B_i} + \mathbf{1}'_{F_i}\boldsymbol{\Sigma}_{F_i}^{-1}\boldsymbol{\Sigma}_{F_iB_i}\boldsymbol{w}_{B_i}\Big)(\boldsymbol{\Sigma}_{F_i}^{-1}\mathbf{1}_{F_i})_i - (\mathbf{1}'_{F_i}\boldsymbol{\Sigma}_{F_i}^{-1}\mathbf{1}_{F_i})\Big(b_i + (\boldsymbol{\Sigma}_{F_i}^{-1}\boldsymbol{\Sigma}_{F_iB_i}\boldsymbol{w}_{B_i})_i\Big)\Big]$$

with

$$C_i = -(\mathbf{1}'_{F_i}\boldsymbol{\Sigma}_{F_i}^{-1}\mathbf{1}_{F_i})(\boldsymbol{\Sigma}_{F_i}^{-1}\boldsymbol{\mu}_{F_i})_i + (\mathbf{1}'_{F_i}\boldsymbol{\Sigma}_{F_i}^{-1}\boldsymbol{\mu}_{F_i})(\boldsymbol{\Sigma}_{F_i}^{-1}\mathbf{1}_{F_i})_i$$

can be rewritten as

$$\lambda^{(i)} = \frac{1}{D_i}\big[\big(1 - \mathbf{1}'_{B}\boldsymbol{w}_{B} + \mathbf{1}'_{F}\boldsymbol{\Sigma}_{F}^{-1}\boldsymbol{\Sigma}_{FB}\boldsymbol{w}_{B}\big)(1 - \boldsymbol{a}'\boldsymbol{\Sigma}_{F}^{-1}\mathbf{1}_{F}) + \big(\boldsymbol{a}'\boldsymbol{\Sigma}_{F}^{-1}\boldsymbol{\Sigma}_{FB}\boldsymbol{w}_{B} - \boldsymbol{\Sigma}_{iB}\boldsymbol{w}_{B}\big)(\mathbf{1}'_{F}\boldsymbol{\Sigma}_{F}^{-1}\mathbf{1}_{F})\big] \tag{12.17}$$

where the used variables are defined as

$$\boldsymbol{\Sigma}_{F_i} = \begin{bmatrix} \boldsymbol{\Sigma}_F & \boldsymbol{a} \\ \boldsymbol{a}' & \alpha \end{bmatrix}, \qquad \boldsymbol{\mu}_{F_i} = \begin{bmatrix} \boldsymbol{\mu}_F \\ \mu_i \end{bmatrix}$$

and

$$D_i = (1 - \boldsymbol{a}'\boldsymbol{\Sigma}_{F}^{-1}\mathbf{1}_{F})(\mathbf{1}'_{F}\boldsymbol{\Sigma}_{F}^{-1}\boldsymbol{\mu}_{F}) - (\mu_i - \boldsymbol{a}'\boldsymbol{\Sigma}_{F}^{-1}\boldsymbol{\mu}_{F})(\mathbf{1}'_{F}\boldsymbol{\Sigma}_{F}^{-1}\mathbf{1}_{F}).$$

Proof. *The derivation is shown in the Appendix.* □

It is remarkable that in (12.17) the vector $\boldsymbol{a}$ (the ith column of Σ corresponding to a trial $i \in \mathbb{B}$) enters linearly. Therefore, the calculation of $\lambda^{(i)}$ is fast and one should calculate only two scalar products with $\boldsymbol{a}$ for each $i \in \mathbb{B}$.

12.4 Performance Tests

Obviously, it is problematic to compare different algorithms based on absolute CPU times. Their performance will strongly depend on the programming language and on the algorithm's memory requirements. When not testing all algorithms on the same computer, different processor performance, RAM size, and system configurations do not allow for a simple interpretation of the differences in CPU run times.

Therefore, the following has to be kept in mind when comparing run times. First, when looking at the increase of CPU time at an increasing number of assets, the algorithms' relative performance is independent from the programming language and other hardware and software properties (as long as there are no memory bottlenecks). Second, the difference in the programming languages does not explain that amount of CPU time improvement such as obtained by our tests.

In the following, a Fortran 90 implementation (Fortran 90 CLA) of the discussed algorithm is tested against three programs for the case where the lower bound is zero and the upper bound is infinity; a simplex-like algorithm based on Wolfe (1959) coded in Java [Java Wolfe-Simplex as described and implemented in Niedermayer (2005)]; and the quadratic optimization package of Matlab. Furthermore, we compare our results with those in Steuer et al. (2006),[13] whose simplex-based multiparametric optimization algorithm was implemented in Java (Java MPQ). The latter comparison is important; as argued in Steuer et al. (2006), the MPQ outperforms Matlab, Cplex, LINGO, Mathematica, and Excel's premium Solver. Steuer et al. (2006) did not compare the Java MPQ algorithm to the Excel Optimizer by Markowitz and Todd (2000) due to the 256 column limitation of Excel. Finally, we also provide run times of the Excel Optimizer by Markowitz and Todd (2000). Note that this implementation is provided by Markowitz and Todd (2000) for illustrative purposes in form of an Excel VBA macro and can calculate the efficient frontier for up to 256 securities (the maximal number of columns in Excel). Note further that even though we ran the Optimizer with the same set of constraints as the other problems, it can solve the optimization problem for a more general set of constraints.

For the tests illustrated in Fig. 12.2, a positive definite covariance matrix was generated as

$$\boldsymbol{\Sigma} = \sum_{i=1}^{n} \boldsymbol{r}^{(i)} \boldsymbol{r}^{(i)\prime},$$

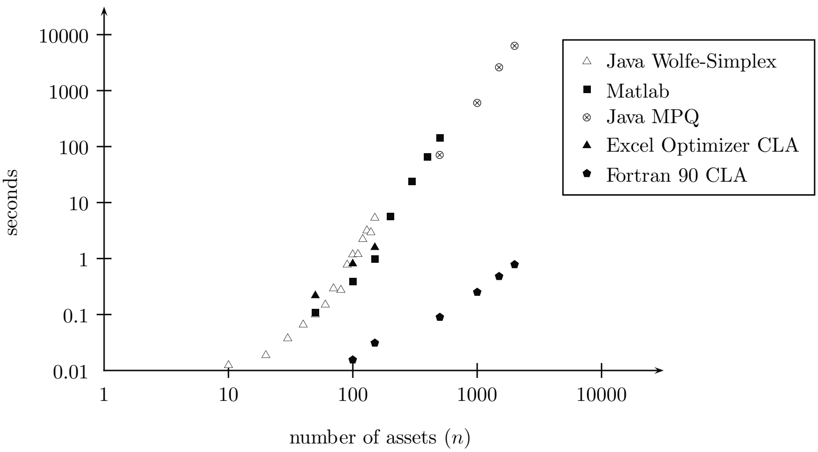


Fig. 12.2 Testing different algorithms for the case with lower bounds zero and upper bounds infinity: CPU times for different number of assets and randomly generated positive definite covariance matrix

[13] Similarly to Steuer et al. (2006), we ran our tests on a Dell 3.06-GHz desktop.

Table 12.1 Different CPU times in seconds for the case with lower bounds zero and upper bounds infinity

n	Fortran 90 CLA	Java Wolfe-Simplex	Matlab	Java MPQ	Excel optimizer CLA
50	–	0.10	0.109	–	0.219
100	0.0156	1.18	0.391	–	0.813
150	0.0312	5.35	0.985	–	1.578
500	0.09	–	141.6	72	–
1,000	0.25	–	–	602	–
1,500	0.48	–	–	2,580	–
2,000	0.78	–	–	6,300	–
Perf.	$\mathcal{O}(n^{1.6})$	$\mathcal{O}(n^{3.6})$	$\mathcal{O}(n^{3.2})$	$\mathcal{O}(n^{3.3})$	$\mathcal{O}(n^{1.8})$

The last row shows the estimates of the algorithms' performance with respect to the number of securities n. Note that the results of the MPQ performance stem from Hirschberger et al. (2004). Note further that the performance we have provided for the Fortran 90 CLA is calculated from the run times without matrix sizes 100 and 150 because for smaller matrix sizes the fixed costs of calculation seem to distort the data. When including matrix sizes 100 and 150 we get $\mathcal{O}(n^{1.3})$

where $\boldsymbol{r}^{(i)}$ is an n vector containing random numbers between [0, 1] and is regenerated for each i. Since our results and the MPQ results in Steuer et al. (2006) strongly depend on the number of free assets (i.e., assets not on their bounds), thus, of the maximum dimension, $\hat{k}$, of $\boldsymbol{\Sigma_F}$ in (12.8) and (12.12), we made sure that $\hat{k}/n$ is similar to that in Steuer et al. (2006).[14] In our tests with 1,000 assets $\hat{k}$ was 60 and when testing with 2,000 assets $\hat{k}$ was 250.

In Fig. 12.2, both axes are logarithmic. The slope of the linear OLS fit corresponds, therefore, to the exponent of the respective algorithm's CPU time increase at an increasing number of assets. Note that the problem with Java Wolfe-Simplex is that the program's RAM requirements increase rapidly which allows only for the computation of problems up to 150 assets. The test's results are summarized in Table 12.1.

Since the Wolfe-Simplex algorithm and the Matlab quadratic optimization package only calculate one single point on the CMVF and do not calculate the whole frontier analytically such as our Fortran CLA algorithm, the Optimizer by Markowitz and Todd (2000), and the algorithm of Steuer et al. (2006), the CPU times reported in Table 12.1 support our method even more.

12.5 Conclusions

This paper presents the CLA developed by Markowitz and demonstrates its strong computational performance compared to standard software packages and to a recently published Optimization algorithm. We find that our implementation of the

[14] Generating $\boldsymbol{r}^{(i)} \sim u[-1/2, 1/2]$ would lead to different values of $\hat{k}/n$.

CLA[15] outperforms the current Matlab optimization tool by a factor of approximately 15,000 when the problem size (number of assets) is 2,000. When comparing with the algorithm in Steuer et al. (2006) that also computes all turning points analytically such as the CLA does, the performance improvement is still around 8,000.

In this paper, we treat all steps of the algorithm explicitly. The algorithm can be used for problems of large-scale portfolio optimization and CPU time-intensive Monte Carlo simulations. The pseudocode we provide is helpful for the implementation of the algorithm in other programming languages.

Appendix

Proofs

Proof (Proposition 12.2). For tractability we define three constants

$$C_{11} \equiv \mathbf{1}'_F \boldsymbol{\Sigma}_F^{-1} \mathbf{1}_F\,, \qquad C_{1\mu} \equiv \mathbf{1}'_F \boldsymbol{\Sigma}_F^{-1} \boldsymbol{\mu}_F\,, \qquad C_{\mu\mu} \equiv \boldsymbol{\mu}'_F \boldsymbol{\Sigma}_F^{-1} \boldsymbol{\mu}_F\,.$$

From (12.4) and the definitions given in (12.1) follows that

$$\boldsymbol{\mu}'\boldsymbol{w} = \boldsymbol{\mu}'_F \boldsymbol{w}_F + \boldsymbol{\mu}'_B \boldsymbol{w}_B = -\boldsymbol{\mu}'_F \boldsymbol{\Sigma}_F^{-1} \boldsymbol{\Sigma}_{FB} \boldsymbol{w}_B + \gamma C_{1\mu} + \lambda C_{\mu\mu}\,.$$

Differentiating this expression with respect to λ yields

$$\frac{\partial(\boldsymbol{\mu}'\boldsymbol{w})}{\partial\lambda} = C_{\mu\mu} - \frac{C_{1\mu}^2}{C_{11}}\,. \tag{12.18}$$

Since between two turning points $\boldsymbol{\Sigma}_F$ does not change, $\mu_{\mathrm{p}}(\lambda) = \boldsymbol{\mu}'\boldsymbol{w}(\lambda)$ is linear in λ with a slope given by (12.18). We show below that this slope is positive.

From the positive definiteness of $\boldsymbol{\Sigma}$ follows that its submatrix $\boldsymbol{\Sigma}_F$ and $\boldsymbol{\Sigma}_F^{-1}$ ($\equiv (\boldsymbol{\Sigma}_F)^{-1}$) are positive definite as well.

We introduce a vector $\boldsymbol{x} \equiv \mathbf{1}_F - \alpha\boldsymbol{\mu}_F$ with $\alpha \in \mathbb{R}$. Then $\boldsymbol{x}'\boldsymbol{\Sigma}_F^{-1}\boldsymbol{x}$ can be written as

$$(\mathbf{1}_F - \alpha\boldsymbol{\mu}_F)'\boldsymbol{\Sigma}_F^{-1}(\mathbf{1}_F - \alpha\boldsymbol{\mu}_F) = C_{11} - 2\alpha C_{1\mu} + \alpha^2 C_{\mu\mu}\,.$$

[15] The program is available on request from the authors in form of a Matlab package at http://www.niedermayer.ch/cla

Positive definiteness of $\boldsymbol{\Sigma}_{\boldsymbol{F}}^{-1}$ means $\boldsymbol{x}'\boldsymbol{\Sigma}_{\boldsymbol{F}}^{-1}\boldsymbol{x} > 0$ for any vector $\boldsymbol{x}$, hence, the equation $C_{11} - 2\alpha C_{1\mu} + \alpha^2 C_{\mu\mu} = 0$ cannot have a solution for α (unless $\boldsymbol{\mu}_{\boldsymbol{F}}$ is parallel to $\boldsymbol{1}_{\boldsymbol{F}}$). Therefore, the discriminant is negative which gives

$$C_{11}C_{\mu\mu} - C_{1\mu}^2 > 0\,.$$

□

Proof (Proposition 12.3). According to Lemma 12.1, $\boldsymbol{\Sigma}_{\boldsymbol{F}_i}^{-1}$ can be expressed in terms of $\boldsymbol{\Sigma}_{\boldsymbol{F}}^{-1}$, $\boldsymbol{a}$, and α. Multiplying $\boldsymbol{\Sigma}_{\boldsymbol{F}_i}^{-1}$ by $\boldsymbol{1}_{\boldsymbol{F}_i}$ and $\boldsymbol{\mu}_{\boldsymbol{F}_i}$, respectively, yields

$$\boldsymbol{\Sigma}_{\boldsymbol{F}_i}^{-1}\boldsymbol{1}_{\boldsymbol{F}_i} = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{F}}^{-1}\boldsymbol{1}_{\boldsymbol{F}} - \beta(1 - \boldsymbol{c}'\boldsymbol{1})\boldsymbol{c} \\ \beta(1 - \boldsymbol{c}'\boldsymbol{1}) \end{bmatrix} \tag{12.19}$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{F}_i}^{-1}\boldsymbol{\mu}_{\boldsymbol{F}_i} = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{F}}^{-1}\boldsymbol{\mu}_{\boldsymbol{F}} - \beta(\mu_i - \boldsymbol{c}'\boldsymbol{\mu}_{\boldsymbol{F}})\boldsymbol{c} \\ \beta(\mu_i - \boldsymbol{c}'\boldsymbol{\mu}_{\boldsymbol{F}}) \end{bmatrix}. \tag{12.20}$$

Multiplying (12.19) and (12.20) by $\boldsymbol{1}'_{\boldsymbol{F}_i}$ and plugging the values into (12.12) yields the denominator in (12.17).

Using the following properties and Lemma 12.1 yields the numerator in (12.17):

$$\begin{aligned} \boldsymbol{1}'_{\boldsymbol{B}_i}\boldsymbol{w}_{\boldsymbol{B}_i} &= \boldsymbol{1}'_{\boldsymbol{B}}\boldsymbol{w}_{\boldsymbol{B}} - b_i\,, \\ \boldsymbol{\Sigma}_{\boldsymbol{F}_i\boldsymbol{B}_i}\boldsymbol{w}_{\boldsymbol{B}_i} &= \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{F}\boldsymbol{B}}\boldsymbol{w}_{\boldsymbol{B}} - \boldsymbol{a}b_i \\ \boldsymbol{\Sigma}_{i\boldsymbol{B}}\boldsymbol{w}_{\boldsymbol{B}} - \alpha b_i \end{bmatrix}. \end{aligned}$$

□

Pseudocode

The following pseudocode describes the procedure which calculates all turning points of the minimum variance frontier for an arbitrary lower bound $\boldsymbol{l}$ and upper bound $\boldsymbol{u}$. Parameters are the expected returns $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. Asset weights $\boldsymbol{w}^{(t)}$ are returned for each turning point t. Between turning points, weights are linear combinations of the two surrounding turning points.

In our notation, $x \leftarrow y$ means that the value y is assigned to the variable x. We define $\arg\max_i\{x_i\}$ to return i^* with $x_{i^*} \geq x_i$ for all i or NIL if the set $\{x_i\}$ is empty. $\max\{\cdot\}$ returns the greatest value unequal to NIL. As in the main text, $\boldsymbol{\Sigma}_{\boldsymbol{F}}$ is a short-hand for the matrix $\{\Sigma_{ij}|i, j \in \mathbb{F}\}$ and $\boldsymbol{\mu}_{\boldsymbol{F}} \equiv \{\mu_i|i \in \mathbb{F}\}$. The same applies to $\boldsymbol{\Sigma}_{\boldsymbol{F}_i}$ and $\boldsymbol{\mu}_{\boldsymbol{F}_i}$ with $\mathbb{F}_i \equiv \mathbb{F} \cup \{i\}$. Similarly, $\mathbb{B}_i \equiv \mathbb{B}\backslash\{i\}$ with $\mathbb{B}$ being defined as the complement of $\mathbb{F}$, i.e., $\mathbb{B} \equiv \{1, \ldots, n\}\backslash\mathbb{F}$. Note that the performance of the algorithm improves significantly if one uses (12.17) from Proposition 12.3 instead of (12.12) on lines 7 and 8 in Procedure ASSET-BECOMES-FREE.

CALCULATE-TURNINGPOINTS($\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{l}, \boldsymbol{u}$)

$\left(\mathbb{F}, \boldsymbol{w}^{(0)}\right) \leftarrow$ STARTING-SOLUTION($\boldsymbol{\mu}, \boldsymbol{l}, \boldsymbol{u}$)
$\lambda_{\text{current}} \leftarrow \infty$
$t \leftarrow 0$
repeat
        ▷ Case a) Free asset moves to its bound
        $(i\_inside, \lambda_{i\_inside}, b) \leftarrow$ ASSET-MOVES-TO-BOUND($\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{l}, \boldsymbol{u}$, $\mathbb{F}, \lambda_{\text{current}}, \boldsymbol{w}^{(t)}$)
        ▷ Case b) Asset on its bound becomes free
        $(i\_outside, \lambda_{i\_outside}) \leftarrow$ ASSET-BECOMES-FREE($\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbb{F}$, $\lambda_{\text{current}}, \boldsymbol{w}^{(t)}$)

  ▷ Find turning point by comparing cases
  if $i\_inside \neq$ NIL or $i\_outside \neq$ NIL
    then $t \leftarrow t+1$
        $\boldsymbol{w}_{\boldsymbol{B}}^{(t)} \leftarrow \boldsymbol{w}_{\boldsymbol{B}}^{(t-1)}$
        $\lambda_{\text{current}} \leftarrow \max\{\lambda_{i\_inside}, \lambda_{i\_outside}\}$
        if $\lambda_{i\_inside} = \max\{\lambda_{i\_inside}, \lambda_{i\_outside}\}$
          then
              $\mathbb{F} \leftarrow \mathbb{F} \backslash \{i\_inside\}$
              $w_{i\_inside}^{(t)} \leftarrow b$
          else
              $\mathbb{F} \leftarrow \mathbb{F} \cup \{i\_outside\}$
        $\gamma \leftarrow -\lambda_{\text{current}} \frac{\boldsymbol{1}_{\boldsymbol{F}}' \boldsymbol{\Sigma}_{\boldsymbol{F}}^{-1} \boldsymbol{\mu}_{\boldsymbol{F}}}{\boldsymbol{1}_{\boldsymbol{F}}' \boldsymbol{\Sigma}_{\boldsymbol{F}}^{-1} \boldsymbol{1}_{\boldsymbol{F}}} + \frac{1 - \boldsymbol{1}_{\boldsymbol{B}}' \boldsymbol{w}_{\boldsymbol{B}}^{(t)} + \boldsymbol{1}_{\boldsymbol{F}}' \boldsymbol{\Sigma}_{\boldsymbol{F}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{FB}} \boldsymbol{w}_{\boldsymbol{B}}^{(t)}}{\boldsymbol{1}_{\boldsymbol{F}}' \boldsymbol{\Sigma}_{\boldsymbol{F}}^{-1} \boldsymbol{1}_{\boldsymbol{F}}}$.
        $\boldsymbol{w}_{\boldsymbol{F}}^{(t)} \leftarrow -\boldsymbol{\Sigma}_{\boldsymbol{F}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{FB}} \boldsymbol{w}_{\boldsymbol{B}}^{(t)} + \gamma \boldsymbol{\Sigma}_{\boldsymbol{F}}^{-1} \boldsymbol{1}_{\boldsymbol{F}} + \lambda_{\text{current}} \boldsymbol{\Sigma}_{\boldsymbol{F}}^{-1} \boldsymbol{\mu}_{\boldsymbol{F}}$
  until $i\_inside =$ NIL and $i\_outside =$ NIL
  ▷ We do not return the starting solution $\boldsymbol{w}^{(0)}$ as it coincides with $\boldsymbol{w}^{(1)}$
  return $\left(\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \ldots, \boldsymbol{w}^{(t)}\right)$

STARTING-SOLUTION($\boldsymbol{\mu}, \boldsymbol{l}, \boldsymbol{u}$)

$\boldsymbol{w} \leftarrow \boldsymbol{l}$
$i \leftarrow \arg\max_j\{\mu_j\}$
while $\boldsymbol{1}'\boldsymbol{w} < 1$
    do
        $i\_free \leftarrow i$
        $w_i \leftarrow \min\{u_i, l_i + 1 - \boldsymbol{1}'\boldsymbol{w}\}$
        $i \leftarrow \arg\max_j\{\mu_j \mid \mu_j < \mu_i\}$
$\mathbb{F} \leftarrow \{i\_free\}$
return $(\mathbb{F}, \boldsymbol{w})$

ASSET-MOVES-TO-BOUND$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{l}, \boldsymbol{u}, \mathbb{F}, \lambda_{\text{current}}, \boldsymbol{w})$

$\triangleright$ A sole free asset cannot move to bound
if *size*$(\mathbb{F}) = 1$
  then
    return (NIL, NIL, NIL)
for $i \in \mathbb{F}$
  do $C_i \leftarrow -(\mathbf{1}'_F \boldsymbol{\Sigma}_F^{-1} \mathbf{1}_F)(\boldsymbol{\Sigma}_F^{-1} \boldsymbol{\mu}_F)_i + (\mathbf{1}'_F \boldsymbol{\Sigma}_F^{-1} \boldsymbol{\mu}_F)(\boldsymbol{\Sigma}_F^{-1} \mathbf{1}_F)_i$
    if $C_i > 0$
      then $b_i \leftarrow u_i$
    if $C_i < 0$
      then $b_i \leftarrow l_i$
    $\lambda_i \leftarrow$ Eq. (12.8)
$i\_inside \leftarrow \arg\max_{i \in \mathbb{F}} \{\lambda_i | \lambda_i < \lambda_{\text{current}}\}$
return $(i\_inside, \lambda_{i\_inside}, b_{i\_inside})$

ASSET-BECOMES-FREE$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{l}, \boldsymbol{u}, \mathbb{F}, \lambda_{\text{current}}, \boldsymbol{w})$

$\triangleright$ Skip procedure if all assets are free
if *size*$(\mathbb{F}) = n$
  then
    return (NIL, NIL)
for $i \in \mathbb{B}$
  do $\mathbb{F}_i \leftarrow \mathbb{F} \cup \{i\}$
    $C_i \leftarrow -(\mathbf{1}'_{F_i} \boldsymbol{\Sigma}_{F_i}^{-1} \mathbf{1}_{F_i})(\boldsymbol{\Sigma}_{F_i}^{-1} \boldsymbol{\mu}_{F_i})_i + (\mathbf{1}'_{F_i} \boldsymbol{\Sigma}_{F_i}^{-1} \boldsymbol{\mu}_{F_i})(\boldsymbol{\Sigma}_{F_i}^{-1} \mathbf{1}_{F_i})_i$
    $\lambda_i \leftarrow$ Eq. (12.12)
$i\_outside \leftarrow \arg\max_{i \in \mathbb{B}} \{\lambda_i | \lambda_i < \lambda_{\text{current}}\}$
return $(i\_outside, \lambda_{i\_outside})$

Acknowledgment The authors thank Ferenc Niedermayer, William F. Sharpe, Ralph E. Steuer, and Heinz Zimmermann for very helpful comments. We further thank G. Peter Todd for providing us the Excel implementation of the Critical Line Algorithm from Markowitz and Todd (2000).

References

Hirschberger, M., Y. Qi, and R. E. Steuer (2004): *Quadratic Parametric Programming for Portfolio Selection with Random Problem Generation and Computational Experience*, Working papers, Terry College of Business, University of Georgia.

Jorion, P. (1992): "Portfolio Optimization in Practice," *Financial Analysts Journal*, 48(1), 68–74.

Markowitz, H. M. (1952): "Portfolio Selection," *Journal of Finance*, 7(1), 77–91.

Markowitz, H. M. (1956): "The Optimization of a Quadratic Function Subject to Linear Constraints," *Naval Research Logistics Quarterly*, III, 111–133.

Markowitz, H. M. (1959): *Portfolio Selection: Efficient Diversification of Investments*. Wiley, New York, and 1991 2nd ed., Basil Blackwell, Cambridge, MA.

Markowitz, H. M., and P. Todd (2000): *Mean–Variance Analysis in Portfolio Choice and Capital Markets*. Frank J. Fabozzi Associates, New Hope, PA.

Markowitz, H., and N. Usmen (2003): "Resampled Frontiers vs Diffuse Bayes: An Experiment," *Journal Of Investment Management*, 1(4), 9–25.

Michaud, R. (1998): *Efficient Asset Management: A Practical Guide to Stock Portfolio Optimization*. Oxford University Press, Oxford.

Niedermayer, D. (2005): *Portfolio Theory and the Cross-sectional Relation between Expected Returns and Betas*, Master's Thesis, Department of Economics, University of Bern.

Scherer, B. (2002): "Portfolio Resampling: Review and Critique," *Financial Analysts Journal*, 58(6), 98–109.

Scherer, B., and D. R. Martin (2006): *Introduction to Modern Portfolio Optimization with NuOPT, S-PLUS and S+Bayes*, 1st edn. Springer, New York.

Steuer, R. E., Y. Qi, and M. Hirschberger (2006): "Portfolio Optimization: New Capabilities and Future Methods," *Zeitschrift für BWL*, 76(2), 199–219.

Wolf, M. (2006): *Resampling vs. Shrinkage for Benchmarked Managers*, IEW – Working Papers iewwp263, Institute for Empirical Research in Economics – IEW, Available at http://ideas.repec.org/p/zur/iewwpx/263.html

Wolfe, P. (1959): "The Simplex Method for Quadratic Programming," *Econometrica*, 27(3), 382–398.

Chapter 13
Factor Models in Portfolio and Asset Pricing Theory

Gregory Connor and Robert A. Korajczyk

13.1 Portfolio Selection

The mean–variance approach of Markowitz (1952, 1959) is an essential underpinning of modern portfolio theory and asset pricing theory. Rational investors who care about expected returns and the variance of their portfolio will hold, subject to any constraints they face, "efficient" portfolios. That is, they hold portfolios with the highest expected return per unit of risk (either variance or standard deviation). Important insights from portfolio selection are that investors rationally diversify (depending on the constraints they face) and that the volatility of an asset (variance or standard deviation) is not, in general, a good indicator of its contribution to the risk of a portfolio of many assets.

At the risk of oversimplifying the analysis of Markowitz, consider the case in which investors have no constraints on short sales.[1] Let ω be a vector of asset positions, with ω_i denoting the fraction of the investor's wealth in asset i (where $i = 1, 2, \ldots, n$). Positive values of ω_i correspond to long positions and negative values to short positions. We require that 100% of the investor's wealth is allocated to assets, that is, $\omega'\iota = 1$, where ι is an n-vector of ones. Letting μ be the n-vector of expected returns on assets and V be the $n \times n$ covariance matrix of asset returns, the portfolio's expected return (μ_{p}), variance (σ_{p}^2), and standard deviation (σ_{p}) are:

$$\mu_{\mathrm{p}} = \omega'\mu,$$

$$\sigma_{\mathrm{p}}^2 = \omega' V \omega,$$

$$\sigma_{\mathrm{p}} = \sqrt{\omega' V \omega}.$$

G. Connor (✉)
Department of Economics, Finance and Accounting,
National University of Ireland, Maynooth County Kildare, Ireland
e-mail: gregory.connor@nuim.ie

R.A. Korajczyk
Kellogg School of Management, Northwestern University,
2001 Sheridan Road, Evanston, IL 60208-2001, USA

[1] Markowitz considered many more complicated problems than the one considered here.

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_13,
© Springer Science+Business Media, LLC 2010

Let $A = \iota' V^{-1} \mu$, $B = \mu' V^{-1} \mu$, $C = \iota' V^{-1} \iota$, and $D = BC - A^2$. The portfolio weights that minimize variance, for a given expected return, μ_{p}, are given by [see Huang and Litzenberger (1988, Section 13.8) or Cochrane (2001, Section 5.2)]:

$$\omega_{\mathrm{MV}} = a + b\mu_{\mathrm{p}},$$

where

$$a = \frac{1}{D}[BV^{-1}\iota - AV^{-1}\mu],$$
$$b = \frac{1}{D}[CV^{-1}\mu - AV^{-1}\iota].$$

Even in this simple case, the solution for minimum-variance portfolio weights requires the inversion of the covariance matrix, V. In practice, of course, we do not observe V, but must estimate it. When there are many assets the solution may be infeasible, due to the computational burden, or unreliable, due to the estimation error in $\widehat{V}$. One approach is to make simplifying assumptions about the structure of V to make the solution easier. One way to impose such restrictions on the covariance matrix is to assume a factor model as the return generating process.

13.2 Basic Definition of a Factor Model

Factor models of security returns decompose the random return on each of a cross section of assets into factor-related and asset-specific returns. Letting r denote the vector of random returns on n assets, and assuming k factors, a factor decomposition has the form

$$r = a_0 + Bf + \varepsilon, \tag{13.1}$$

where B is a $n \times k$ matrix of factor betas, f is a random k-vector of factor returns, and ε is an n-vector of asset-specific returns. The n-vector of coefficients a_0 is set so that $E[\varepsilon] = 0$. By defining B as the least squares projection $B = \mathrm{cov}(r, f)V_f^{-1}$, it follows that $\mathrm{cov}(f, \varepsilon) = 0^{k\times n}$.

The factor decomposition (13.1) puts no empirical restrictions on returns beyond requiring that the means and variances of r and f exist. So in this sense it is empty of empirical content. To add empirical structure, it is commonly assumed that the asset-specific returns ε are cross-sectionally uncorrelated, $E[\varepsilon\varepsilon'] = D$, where D is a diagonal matrix. This implies that the covariance matrix of returns can be written as the sum of a matrix of rank k and a diagonal matrix:

$$V = \mathrm{cov}(r, r') = BV_f B' + D. \tag{13.2}$$

This is called a strict factor model. Without loss of generality one can assume that $V_f = \text{cov}(f, f')$ has rank k, since otherwise one of the factors can be removed (giving a $k-1$ factor model) without affecting the fit of the model.

If the vector of factors, f, is a specified set of random variables (e.g., the returns a set of market indices), we will call the model a prespecified, or macroeconomic, factor model. If the vector of factors, f, is unknown, and must be estimated, we will call the model a statistical factor model. If the matrix of factor loadings, B, is known, but f must be estimated, we will call the model a characteristic factor model (Connor 1995). There are also applications which involve hybrid factor models: ones which have some prespecified factors and/or loadings and some statistical factors (Stroyny 2005).

Estimation of the unrestricted covariance matrix of n securities, V, requires the estimation of $n \times (n+1)/2$ distinct elements. The single index, or diagonal, model of Sharpe (1963) is a single-factor prespecified strict factor model. It assumes that all of the common elements of returns were due to assets' relations with the market index. Thus, only $(3 \times n) + 1$ parameters needed to be estimated: n elements of B, or "betas" relative to the index, n unique variances, n intercept terms, and the index variance, σ_f^2. This approach reduces much of the noise in the estimate of V. In practice, the single index, diagonal model does not describe all of the common movements across assets (i.e., the residual matrix is not diagonal) so there seems to be some additional benefit from using a multifactor model. With k factors there are still only $n \times (k+2) + k \times (k+1)/2$ parameters to estimate [$n \times k$ betas, n intercepts or means, n unique variances, and $k \times (k+1)/2$ elements of the factor covariance matrix]. An alternative approach is to continue with a smaller set of factors, but place other restrictions on the covariance matrix of ε. Some early studies in this area are Farrar (1962), King (1966), Cohen and Pogue (1967), and Elton and Gruber (1973).

A strict factor model aids in computing optimal portfolio weights since

$$V^{-1} = D^{-1} - D^{-1} B V_f (V_f + V_f B' D^{-1} B V_f)^{-1} V_f B' D^{-1}$$

[see Muirhead (1982, p. 580)]. The strict factor model only requires the inversion of an $n \times n$ diagonal matrix, D, and a $k \times k$ matrix, $V_f + V_f B' D^{-1} B V_f$. Since k is typically very small, relative to the number of assets, n, these inverses are easy to calculate and are likely to be more stable than the inverse of an unrestricted $n \times n$ covariance matrix. The factor model trades off the potential bias that the residual covariance matrix may not be diagonal with the added precision of the model estimates due to the higher numbers of observations per parameter.

13.3 Statistical Factor Models

We differentiate between characteristic-based, macroeconomic, and statistical factor models. In a characteristic-based model, the factor betas of asset are tied to observable characteristics of the securities, such as company size or the book-to-price ratio, or the industry categories to which each security belongs. In macroeconomic factor models, the factors are linked to the innovations in observable economic time series such as inflation and unemployment. In a statistical factor model, neither factors nor betas are tied to any external data sources and the model is identified from the covariances of asset returns alone.[2]

The convenient rotation $E[ff'] = I$ allows us to write the strict factor model (13.1) as:

$$\operatorname{cov}(r, r') = BB' + D. \tag{13.3}$$

Assuming that the cross section of return is multivariate normal and i.i.d. through time, the sample covariance matrix $\widehat{\operatorname{cov}}(r, r')$ has a Wishart distribution. Imposing the strict factor model assumption (13.3) on the true covariance matrix, it is possible to estimate the set of parameters B, D by maximum likelihood. This maximum likelihood problem requires high-dimensional nonlinear maximization: there are $nK + n$ parameters to estimate in B, D. There is also an inequality constraint on the maximization problem: the diagonal elements of D must be nonnegative, since they represent variances. The solution to the maximum likelihood problem yields estimates of B and D which correspond to the systematic and unsystematic risk measures. It is often the case that estimates of the time series of factors, f, are of interest. These are called factor scores in the statistical literature and can be obtained through cross-sectional GLS regressions of r on B

$$\widehat{f}_t = (\widehat{B}'\widehat{D}^{-1}\widehat{B})^{-1}\widehat{B}'\widehat{D}^{-1}r_t.$$

See Basilevsky (1994) for a review of the various iterative algorithms which can be used to numerically solve the maximum likelihood factor analysis problem and estimate the factor scores. See Roll and Ross (1980) for an empirical application to equity returns data.

The first k eigenvectors of the return covariance matrix scaled by the square roots of their respective eigenvalues are called the k *principal components* of the covariance matrix. A restrictive version of the strict factor model is the *scalar factor model*, given by (13.2) plus the scalar matrix condition $D = \sigma_\varepsilon^2 I$. Under the assumption of a scalar factor model, the maximum likelihood problem simplifies, and the principal components are the maximum likelihood estimates of the factor beta matrix B (the arbitrary choice of rotation is slightly different in this case). This provides a quick and simple alternative to maximum likelihood factor analysis, under the restrictive assumption $D = \sigma_\varepsilon^2 I$.

An important (and often troublesome) feature of statistical factor models is their rotational indeterminacy. Let L denote any nonsingular $k \times k$-matrix and consider the set of factors $f^* = Lf$ and factor betas $B^* = BL^{-1}$. Note that f^*, B^* can be used in place of f, B since only their matrix product affects returns and the

[2] See Connor (1995) and Connor and Korajczyk (2009).

linear "rotation" L disappears from this product. This means that factors f and associated factor betas B are only defined up to a $k \times k$ linear transformation. In order to empirically identify the factor model, one can set the covariance matrix of the factors equal to an identity matrix, $E[ff'] = I_k$, without loss of generality.

13.3.1 Approximate Factor Models

The assumption that returns obey a strict factor model is easily rejected. In practice, for most reasonable values of k, there will at least some discernible positive correlations between the asset-specific returns of at least some assets. An *approximate factor model* (originally developed by Chamberlain and Rothschild (1983)) weakens the strict factor model of exactly zero correlations between all asset-specific returns. Instead it assumes that there is a large number of assets n and the proportion of the correlations which are nonnegligibly different from zero is close to zero. This condition is formalized as a bound on the eigenvalues of the asset-specific return covariance matrix:

$$\lim_{n\to\infty} \max \text{eigval}[\text{cov}(\varepsilon, \varepsilon')] < c$$

for some fixed $c < \infty$. Crucially, this condition implies that asset-specific returns are *diversifiable risk* in the sense that any well-spread portfolio w will have asset-specific variance near zero:

$$\lim_{n\to\infty} w'\text{cov}(\varepsilon, \varepsilon')w = 0 \text{ for any } w \text{ such that } \lim_{n\to\infty} w'w = 0. \tag{13.4}$$

Note that an approximate factor model uses a "large n" modeling approach: the restrictions on the covariance matrix need only hold approximately as the number of assets n grows large.

Letting $V_\varepsilon = \text{cov}(\varepsilon, \varepsilon')$ which is no longer diagonal, and choosing the rotation so that $\text{cov}(f, f') = I$ we can write the covariance matrix of returns as:

$$\text{cov}(r, r') = BB' + V_\varepsilon.$$

In addition to (13.4) it is appropriate to impose the condition that $\lim_{n\to\infty} \min BB' = \infty$. This ensures that each of the k factors represents a pervasive source of risk in the cross section of returns.

13.3.2 Asymptotic Principal Components

The maximum likelihood method of factor model estimation relies on a strict factor model assumption and a time-series sample which is large relative to the number of assets in the cross section. Standard principal components require the even stronger condition of a scalar factor model. Neither method is well configured for asset returns where the cross section tends to be very large. Connor and Korajczyk (1986) develop an alternative method called asymptotic principal components, building on the approximate factor model theory of Chamberlain and Rothschild (1983). Connor and Korajczyk analyze the eigenvector decomposition of the $T \times T$ cross-product matrix of returns rather than of the $n \times n$ covariance matrix of returns. They show that given a large cross section, the first k eigenvectors of this cross-product matrix provide consistent estimates of the $k \times T$ matrix of factor returns. Stock and Watson (2002) extend the theory to allow both large time series and large cross-sectional samples, time varying factor betas, and provide a quasi-maximum likelihood interpretation of the technique. Bai (2003) analyzes the large-sample distributions of the factor returns and factor beta matrix estimates in a generalized version of this approach.

13.4 Macroeconomic Factor Models

The rotational indeterminacy in statistical factor models is unsatisfying for the application of factor models to many research problems. Statistical factor models do not allow the analyst to assign meaningful labels to the factors and betas; one can identify the k pervasive risks in the cross section of returns, but not what these risks represent in terms of economic and financial theory.

One approach to making the factor decomposition more interpretable is to rotate the statistical factors so that the rotated factors are maximally correlated with prespecified macroeconomic factors. If f_t is a k-vector of statistical factors and m_t is a k-vector of macroeconomic innovations we can regress the macroeconomic factors on the statistical factors

$$m_t = \Pi f_t + \eta_t.$$

As long as Π has rank k, the span of the rotated factors, Πf_t, is the span of the original statistical factors, f_t. However, the rotated factors can now be interpreted as the return factors that are correlated with the specified macroeconomic series. With this rotation the new factors are no longer orthogonal, in general. This approach is described in Connor and Korajczyk (1991).

Alternatively, one can work with the prespecified macroeconomic series directly. Chan et al. (1985) and Chen et al. (1986) develop a macroeconomic factor model in which the factor innovations f are observed directly (using innovations in economic time series) and the factor betas are estimated via time-series regression of each asset's return on the time series of factors. They begin with the standard description of the current price of each asset, p_{it}, as the present discounted value of its expected cash flows:

$$p_{it} = \sum_{s=1}^{\infty} \frac{E[c_{it}]}{(1 + \rho_{st})^s},$$

where ρ_{st} is the discount rate at time t for expected cash flows at time $t+s$. Chen, Roll, and Ross note that the common factors in returns must be variables which cause pervasive shocks to expected cash flows $E[c_{it}]$ and/or risk-adjusted discount rates ρ_{st}. They propose inflation, interest rate, and business-cycle related variates to capture these common factors. Shanken and Weinstein (2006) find that empirically the model lacks robustness in that small changes in the included factors or the sample period have large effects on the estimates. Connor (1995) argues that although macroeconomic factors models are theoretically attractive since they provide a deeper explanation for return comovement than statistical factor models, their empirical fit is substantially weaker than statistical- and characteristic-based models. Vassalou (2003) argues on the other hand that the ability of the Fama-French model (see below) to explain the cross section of mean returns can be attributed to the fact that Fama-French factors provide good proxies for macroeconomic factors.

13.5 Characteristic-Based Factor Models

A surprisingly powerful method for factor modeling of security returns is the characteristic-based factor model. Rosenberg (1974) was the first to suggest that suitably scaled versions of standard accounting ratios (book-to-price ratio, market value of equity) could serve as factor betas. Using these predefined betas, he estimates the factor realizations f_t by cross-sectional regression of time-t asset returns on the predefined matrix of betas.

In a series of very influential papers, Fama and French (1992, 1993, 1996) propose a two-stage method for estimating characteristic-based factor models. In the first stage, they sort assets into portfolios based on book-to-price and market-value characteristics. They use the differences between returns on the top and bottom fractile portfolios as proxies for the factor returns. They also include a market factor proxied by the return on a capitalization-weighted market index. In the second stage, the factor betas of portfolios and/or assets are estimated by time-series regression of asset returns on the derived factors. Carhart (1997) and Jegadeesh and Titman (1993, 2001) show that the addition of a momentum factor (proxied by high-12-month return minus low-12-month return) adds explanatory power to the Fama-French three-factor model, both in terms of explaining comovements and mean returns. Ang etal. (2006, 2009) and Goyal and Santa-Clara (2003) also find evidence for a own-volatility-related factor, both for explaining return comovements and mean returns.

One of the most empirically powerful factor decompositions for equity returns is an error-components model using industry affiliations. This involves setting the factor beta matrix equal to zero/one dummies, with row i containing a one in the jth column if and only if firm i belongs to industry j. This is the simplest type of characteristic-based factor model of equity returns.

The first statistical factor is dominant in equity returns, accounting for 80–90% of the explanatory power in a multifactor model. The standard specification of a error-components model does not isolate the "first" factor since its influence is spread across the factors. Heston and Rouwenhorst (1994) describe an alternative specification in which this factor is separated from the k industry factors. They add a constant to the model, so that the expanded set of factors $k + 1$ is not directly identified (this lack of identification is sometimes called the "dummy variable trap," referring to a model that includes a full set of zero-one dummies plus a constant). Then, Heston and Rouwenhorst impose an adding-up restriction on the estimated $k+1$ factors: the set of industry factors must sum to zero. This adding-up restriction on the factors restores statistical identification to the model, requiring constrained least squared in place of standard least squares estimation. It also provides a useful interpretation of the estimated factors: the factor associated with the constant term is the "market-wide" or "first" factor, and the factors associated with the industry dummies are the extra-market industry factors.

Heston and Rouwenhorst's adding-up condition is particularly useful in a multi-country context. It allows one to include an overlapping set of country and industry dummies without encountering the problem of the dummy variable trap. Including a constant, an international industry–country factor model must impose adding-up conditions both on the estimated industry factors and on the estimated country factors. This type of country–industry specification is useful for example in measuring the relative contribution of cross-border and national influences to return comovements, see, for example, Hopkins and Miller (2001).

13.6 Mutual Fund Separation

Given a set of traded assets and investors with different tastes for risk, it might be the case that each investor's portfolio is customized to the preferences of that investor and is quite different in composition than another investor's portfolio. When individual investors' optimal portfolio choices across all n assets can be reduced to the choice of combining $k < n$ portfolios, we say that k-fund separation holds. With k-fund separation, each investor is indifferent between choosing from the full set of n assets or from the set of k mutual funds. Tobin (1958) proves a one-fund separation theorem for investors choosing between cash and an array of money market instruments.

Separation results can be obtained either for particular specifications of agents' utility functions or for certain specifications of return distributions. Cass and Stilglitz (1970) derive necessary and sufficient conditions on utility functions to get separation. Given our emphasis of the role of factor models in portfolio selection and asset pricing, we will focus on the relation between return distributions and mutual fund separation. Ross (1978b) studies the conditions on return distributions that lead to mutual fund separation. Let x denote the payoff vector of a set of n assets next period. Weak k-fund separation holds if there exist k

mutual funds, $\alpha^1, \alpha^2, \ldots, \alpha^k$, such that, for any portfolio, θ, and any monotonically increasing and concave (utility) function, u, there exists a portfolio α

$$\alpha = a_1\alpha^1 + a_2\alpha^2 + \cdots + a_k\alpha^k$$

with

$$E\{u(\alpha x)\} \geq E\{u(\theta x)\}.$$

Ross (1978b, Theorem 3) shows that necessary and sufficient conditions for weak k-fund separation are that returns follow a k-factor data generating process, that expected returns on assets are linearly dependent on their factor exposures, and that the k mutual funds have no nonfactor risk. Thus, k-fund separation requires a restriction on the vector of mean returns in addition to restrictions on the covariance matrix of returns. The restriction on expected returns is inherently an asset pricing question, which we discuss in Sect. 13.7.

13.7 Asset Pricing Theory

Mean–variance analysis is a cornerstone of Asset Pricing Theory. Let x denote the payoff vector of a set of n assets next period, $p(x)$ denote the price of x and X denote the set of all payoffs that an investor can purchase. Assume that markets are competitive and frictionless, in the sense that there are no transactions costs, indivisibilities, or short sale constraints. The law of one price is the condition that two identical payoffs must have the same price. Thus, if payoff y is the same as a units of payoff x and b units of payoff z, then $p(y) = a \times p(x) + b \times p(z)$. Given the assumption of competitive and frictionless markets and the assumption of the law of one price (linearity of the pricing function), then there exists a unique payoff $x^* \in X$ for which the price of any payoff x in X is given by:

$$p(x) = E(x^* x) \tag{13.5}$$

(see Cochrane (2001, 4.1) for a proof). The random variable x^* is often referred to as a discount factor, pricing kernel, or state-price density. There may be other discount factors that are not in X which also price assets. In fact, any discount factor, m, with $m = x^* + \nu$ and $E(\nu x) = 0$ will also price all payoffs in X since $E(mx) = E(x^* x) + E(\nu x) = p(x) + 0 = p(x)$.

Define an arbitrage opportunity as the existence of a payoff, x, that is nonnegative ($x \geq 0$), that is positive with positive probability (i.e., $\Pr(x > 0) > 0$), and that which has a nonpositive price ($p(x) \leq 0$). In a market where agents prefer more to less, we should not expect arbitrage opportunities to exist. The absence of arbitrage opportunities implies that the discount factor is strictly positive [see Ross (1978a) and Cochrane (2001)].

Given a valid traded discount factor, $m = x^*$, if a riskless asset exists with rate of return equal to r_f and gross payoff $R_f = 1 + r_f$, then $p(R_f) = 1 = E(mR_f) = E(m)R_f$. Therefore,

$$R_f = \frac{1}{E(m)}.$$

Since $E(mx) = E(m)E(x) + \text{cov}(m, x)$, we have that:

$$p(x) = \frac{E(x)}{R_f} + \text{cov}(m, x). \tag{13.6}$$

The price of payoff x is its expected value discounted at the riskless rate plus a risk premium or discount, $\text{cov}(m, x)$. Denote the gross (net) return on x as $R_x(r_x)$. $R_x = \frac{x}{p(x)}$ and $r_x = \frac{x}{p(x)} - 1$. By dividing both sides of (13.6) by $p(x)$ and rearranging, we can express the pricing relation as:

$$E(r_x) - r_f = E(R_x) - R_f = -R_f \times \text{cov}(m, r_x). \tag{13.7}$$

From (13.6) and (13.7) we see that the only risk which influences asset prices (or equivalently expected returns) is the component of asset returns that is correlated with the discount factor, m. Asset expected returns can be expressed in terms of their "betas" relative to the discount factor by multiplying and dividing (13.7) by $\text{var}(m)$ and rearranging to get

$$E(r_x) - r_f = E(R_x) - R_f = \beta_{x,m}\lambda_m, \tag{13.8}$$

where

$$\beta_{x,m} = \frac{\text{cov}(r_x, m)}{\text{var}(m)},$$
$$\lambda_m = \frac{-\text{var}(m)}{E(m)}.$$

Rearranging (13.8) we find that

$$E(R_x) - R_f = \rho_{x,m}\frac{\sigma(m)}{E(m)}\sigma(R_x),$$

where $\rho_{x,m}$ is the correlation between asset returns and the discount factor.

Since $\rho_{x,m}$ is between -1 and 1, we have the following bound on excess returns

$$|E(R_x) - R_f| \leq \frac{\sigma(m)}{E(m)}\sigma(R_x).$$

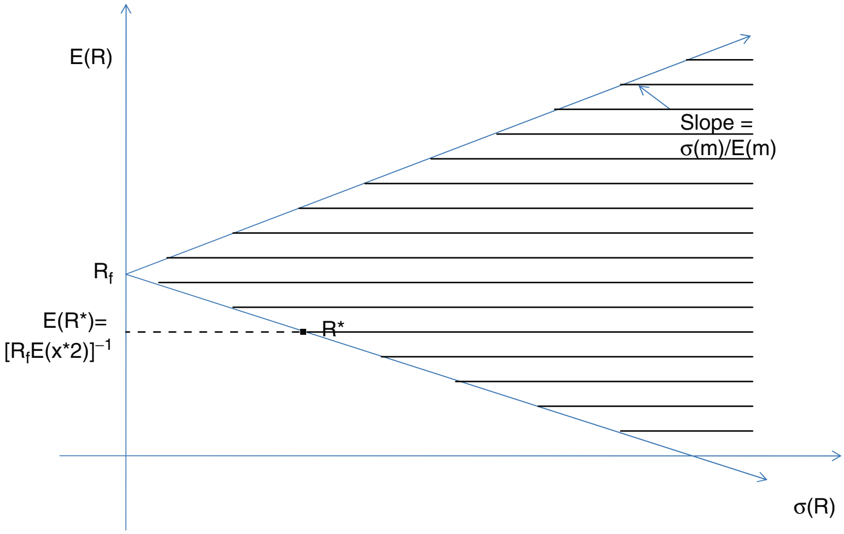


Fig. 13.1 The *cross-hatched area* represents the set of attainable expected return/standard deviation combinations. The mean–variance boundary can be obtained by combinations of $R*$ and R_f

All assets must lie on, or to the right of, the two rays emanating from R_f as shown in the cross-hatched area in Fig. 13.1. These rays are called the mean–variance boundary since they represent the lowest standard deviation possible for a given level of expected return. The upward-sloping portion of the mean–variance boundary is often call the efficient frontier since these portfolios have the highest expected return for a given level of standard deviation. Any portfolio on the mean–variance boundary is perfectly correlated with the discount factor. Let R^* denote a frontier portfolio, $R^* = x^*/p(x^*)$. We can write the discount factor as a linear function of R^*.

So far, we have invoked minimal assumptions to derive (13.8). In particular, we have not assumed specific forms of agents' utility functions or properties of the distributions of asset returns other than the existence of second moments. We have assumed frictionless, competitive markets and the lack of arbitrage opportunities. While identifying portfolios on the mean–variance boundary, ex post, is relatively straightforward, doing so *ex ante* is not. Much of the asset pricing literature is devoted to placing enough structure on the economic problem to derive the identity of m. We will begin by discussing consumption-based models and then factor models of returns.

Assume that agents wish to maximize expected lifetime utility of the following sort

$$E_t \sum_{j=0}^{\infty} \delta^j u(c_{t+1}) \tag{13.9}$$

and that an asset is a claim to a stream of future payments, $d_{t+1}, d_{t+2}, \ldots$ The first-order conditions of the maximization problem imply that the price of the asset is given by

$$p_t = E_t \sum_{j=1}^{\infty} \delta^j \frac{u'(c_{t+1})}{u'(c_t)} d_{t+1}$$

[see Lucas (1978), Brock (1982), and Cochrane (2001)]. This can be rewritten as

$$p_t = E_t \left[\delta \frac{u'(c_{t+1})}{u'(c_t)} (d_{t+1} + p_{t+1}) \right] = E_t \left[m_{t+1} x_{t+1} \right]. \tag{13.10}$$

Therefore, the consumption-based discount factor is $m_{t+1} = \delta \frac{u'(c_{t+1})}{u'(c_t)}$ which prices the payoff to holding the asset one period, $x_{t+1} = d_{t+1} + p_{t+1}$. Assets whose payoffs have higher covariance with future marginal utility (i.e., those that pay well in "bad times" when consumption is low and marginal utility is high) have higher values or, equivalently, lower expected returns. Dividing both sides of (13.10) by p_t yields the moment condition

$$1 = E_t \left[\delta \frac{u'(c_{t+1})}{u'(c_t)} \frac{(d_{t+1} + p_{t+1})}{p_t} \right] = E_t \left[\delta \frac{u'(c_{t+1})}{u'(c_t)} R_{x,t+1} \right], \tag{13.11}$$

which implies

$$E_t \left[\delta \frac{u'(c_{t+1})}{u'(c_t)} R_{x,t+1} - 1 \right] = 0. \tag{13.12}$$

The moment condition in (13.12) can be multiplied by any instrument in the time t information set, say z_t, to get additional moment restrictions

$$E_t \left[\left(\delta \frac{u'(c_{t+1})}{u'(c_t)} R_{x,t+1} - 1 \right) \times z_t \right] = 0. \tag{13.13}$$

Hansen and Singleton (1982) provide one of the earliest empirical tests of the restrictions in equation (13.13) assuming constant relative risk aversion, $u(c) = \frac{c^\gamma}{\gamma}$ for $\gamma < 1$. When equity and bond returns are included, the restrictions implied by the pricing model are rejected. The low correlation between changes in consumption and stock returns requires very high levels of risk aversion to match the return on stocks. With expected utility, the high level of risk aversion implies a low level of intertemporal substitution which implies interest rates much higher than that we observe in the data [see Hansen and Singleton (1982), Mehra and Prescott (1985), Kocherlakota (1996), and Cochrane (2005)]. The failure of the early tests of consumption-based asset pricing models might be due to properties of the consumption data, such as temporal aggregation and measurement error

(Breeden et al. 1989). Alternatively, the specifications of the utility function may be inappropriate. Nonseparable utility specifications, such as habit formation, where utility is defined relative to a reference level of utility based on past consumption, seem to have some success in explaining average returns. Epstein and Zin (1989, 1991) derive and test a model in which utility is not state-separable. Their utility specification separates the level of risk aversion from the intertemporal elasticity of substitution, and hence can fit the empirical regularities of asset returns better than the time- and state-separable specifications. Incorporating shocks to labor or entrepreneurial income, whether tradable [see Jagannathan and Wang (1996) and Campbell (1996)] or nontradable [see Heaton and Lucas (1996, 1999) and Brav et al. (2002)) help in matching mean returns.

Consumption-based asset pricing provides strong economic intuition about the determinants of risk premia. However, the challenges of working with consumption data are formidable. An alternative is suggested by the pricing function, (13.5), and the linear "beta" relation, (13.8). The pricing function shows that there is a portfolio whose payoff can serve as the discount factor. This portfolio is on the mean–variance boundary in Fig. 13.1. In fact, any portfolio on the boundary will work. In terms of the consumption-based discount factor, say $m_{t+1} = \delta\frac{u'(c_{t+1})}{u'(c_t)}$, the portfolio of assets whose payoffs are maximally correlated with $\delta\frac{u'(c_{t+1})}{u'(c_t)}$ will price all assets. Thus, $x^*$is the projection of $\delta\frac{u'(c_{t+1})}{u'(c_t)}$ onto the space of payoffs, X.

The earliest asset pricing literature began by making sufficient assumptions on preferences and/or the distribution of payoffs to identify x^*. With quadratic preferences or joint multivariate normality of returns, myopic preferences, and the lack of other sources of income besides the return on agents' portfolios, we obtain one-fund separation. Given that all agents hold the same mutual fund of risky assets, equilibrium between supply and demand for assets requires that portfolio be the value-weighted market portfolio. This yields the Capital Asset Pricing Model (CAPM) of Sharpe (1964), Lintner (1966), Treynor (1961, 1999), and Mossin (1966). Let r_{M} denote the rate of return on the market portfolio. The CAPM implies that $x^* = a + br_{\mathrm{M}}$ and that

$$E(r_x) - r_f = E(R_x) - R_f = \beta_{x,\mathrm{M}}\lambda_{\mathrm{M}} \tag{13.14}$$

where

$$\beta_{x,\mathrm{M}} = \frac{\mathrm{cov}(r_x, r_{\mathrm{M}})}{\mathrm{var}(r_{\mathrm{M}})}$$

$$\lambda_{\mathrm{M}} = E(r_{\mathrm{M}}) - r_f.$$

The Arbitrage Pricing Theory (APT) of Ross (1976) assumes that asset returns follow a factor model, as in equation (13.1). If the factor model is a noiseless factor model (i.e., $\varepsilon \equiv 0$ in (13.1)), then the law of one price implies that expected returns are linear in assets exposure to the underlying factors. Equivalently the discount factor is a linear function of the factors:

$$\begin{aligned} x^* &= a + b'f, \\ E(r_x) - r_f &= \beta'_{x,f}\lambda_f, \end{aligned} \tag{13.15}$$

where $\beta_{x,f} = \mathrm{cov}(r_x, f)$, a $k \times 1$ vector of covariances of r_x with the k factors and $\lambda_\mathrm{f} = -R_\mathrm{f}\mathrm{cov}(ff')b = -R_f E(x^* f)$.[3]

If the factor model is not noiseless, in the sense that ε is a mean-zero random vector, then we need sufficient conditions for the k mutual funds $\alpha^1, \alpha^2, \ldots, \alpha^k$, discussed in Sect. 13.6, to be well diversified:

$$\alpha^j \varepsilon = 0$$

for $j = 1, 2, \ldots, k$. Connor (1984) derives an equilibrium version of the APT in which there are an infinite number of assets. In this economy we get k-fund separation and multifactor pricing, as in (13.15) if the market portfolio is well diversified.

With a strict factor model and a finite set of assets, we would, in general, not be able to construct k well-diversified mutual funds, so the APT pricing relation, (13.15) will hold as an approximation as derived in Ross (1976).

Alternative multifactor asset pricing models include the Intertemporal Capital Asset Pricing Model (ICAPM) of Merton (1973) in which market risk and the risks associated with shifts in the consumption/investment opportunity set command risk premia. Empirical implementation of macroeconomic factor models often are built by postulating a set of variables that are good proxies for such risks. Examples are Chan et al. (1985), Chen et al. (1986), Shanken (1990), Fama and French (1993, 1996), and Carhart (1997). More recently there has been increased interest in estimating models that include factors that proxy for shocks to liquidity [e.g., Pástor and Stambaugh (2003), Acharya and Pedersen (2005), and Korajczyk and Sadka (2008)]. However, these models step away from the frictionless markets assumption underlying the derivation of the discount factor.

These empirical implementations of factor models essentially take a stand on the set of macroeconomic variables that drive a traded discount factor x^* or a potentially nontraded discount factor $m = x^* + \nu$.

13.8 Summary

The mean–variance analysis of Harry Markowitz provides the basis for portfolio and asset pricing theory. Within those large literatures, factor models have played a significant role. In portfolio theory, strict factor models impose structure on covariance

[3] See Cochrane (2001, Section 6.3).

matrices in ways that make portfolio optimization feasible in situations where unconstrained problems are infeasible or computationally burdensome.

The role of mean–variance boundary portfolios is central in asset pricing. With competitive and frictionless markets and the existence of second moments, the law of one price implies the existence of a traded portfolio whose payoff can be used to price all assets. The portfolio is on the mean–variance boundary (the set of portfolios with minimum variance for each level of expected return). This portfolio payoff will price all assets, including derivative securities, regardless of the distribution of asset returns or the preferences of investors (again assuming preferences rule out arbitrage opportunities and the relevant moments exist).

Empirical applications of asset pricing theory, of course, require the identification of the appropriate discount factor. While consumption-based models provide a great deal of economic intuition, the extant consumption data often are poor proxies for the quantities needed to apply the models. Factor models, whether statistical, macroeconomic, or characteristic based, provide different paths to creating sets of portfolios that provide candidate discount factors.

Acknowledgment Korajczyk wishes to acknowledge the financial support of the Zell Center for Risk Research and the Jerome Kenney Fund.

References

Acharya, Viral V., and Lasse Heje Pedersen, 2005, "Asset Pricing with Liquidity Risk," *Journal of Financial Economics* 77, 375–410.

Ang, Andrew, Robert J. Hodrick, Yuhang Xing, and Xiaoyan Zhang, 2006, "The Cross-Section of Volatility and Expected Returns," *Journal of Finance* 61, 259–299.

Ang, Andrew, Robert J. Hodrick, Yuhang Xing, and Xiaoyan Zhang, 2009, "High Idiosyncratic Risk and Low Returns: International and Further U. S. Evidence," *Journal of Financial Economics* 91, 1–23.

Bai, Jushan, 2003, "Inferential Theory for Factor Models of Large Dimension," *Econometrica* 71, 135–171.

Bai, Jushan, and Serena Ng, 2002, "Determining the Number of Factors in Approximate Factor Models," *Econometrica* 70, 191–221.

Basilevsky, Alexander, 1994, *Statistical Factor Analysis and Related Methods: Theory and Applications*. New York: Wiley.

Brav, Alon, George M. Constantinides, and Christopher C. Geczy, 2002, "Asset Pricing with Heterogeneous Consumers and Limited Participation: Empirical Evidence," *Journal of Political Economy* 110, 793–824.

Breeden, Douglas T., Michael R. Gibbons, and Robert H. Litzenberger, 1989, "Empirical Test of the Consumption-Oriented CAPM," *Journal of Finance* 44, 231–262.

Brock, William A., 1982, "Asset Prices in a Production Economy," in *The Economics of Information and Uncertainty*, edited by J. J. McCall. Chicago, IL: University of Chicago Press.

Campbell, John Y., 1996, "Understanding Risk and Return," *Journal of Political Economy* 104, 298–345.

Campbell, John Y., 2003, "Consumption-Based Asset Pricing," in *Handbook of the Economics of Finance*, edited by G. M. Constantinides, M. Harris, and R. Stulz. Amsterdam: North Holland.

Carhart, Mark M., 1997, "On Persistence in Mutual Fund Performance," *Journal of Finance* 52, 57–82.

Cass, David, and Joseph E. Stiglitz, 1970, "The Structure of Investor Preferences and Asset Returns, and Separability in Portfolio Allocation: A Contribution to the Pure Theory of Mutual Funds," *Journal of Economic Theory* 2, 122–160.

Chamberlain, Gary, and Michael Rothschild, 1983, "Arbitrage, Factor Structure and Mean–Variance Analysis in Large Asset Markets," *Econometrica* 51, 1305–1324.

Chan, K. C., Nai-fu Chen, and David Hsieh, 1985, "An Exploratory Investigation of the Firm Size Effect." *Journal of Financial Economics* 14, 451–471.

Chen, Nai-Fu, Richard Roll, and Stephen A. Ross, 1986, "Economic Forces and the Stock Market," *Journal of Business* 59, 383–403.

Cochrane, John H., 2001, *Asset Pricing*, Princeton, NJ: Princeton University Press.

Cochrane, John H., 2005, "Financial Markets and the Real Economy," *Foundations and Trends in Finance* 1, 1–101.

Cohen, Kalman J., and Jerry A. Pogue, 1967, "An Empirical Evaluation of Alternative Portfolio-Selection Models," *Journal of Business* 40, 166–193.

Connor, Gregory, 1984, "A Unified Beta Pricing Theory," *Journal of Economic Theory* 34, 13–31.

Connor, Gregory, 1995, "The Three Types of Factor Models: A Comparison of Their Explanatory Power," *Financial Analysts Journal* 51, 42–46.

Connor, Gregory, and Robert A. Korajczyk, 1986, "Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis," *Journal of Financial Economics* 15, 373–394.

Connor, Gregory, and Robert A. Korajczyk, 1991, "The Attributes, Behavior, and Performance of U.S. Mutual Funds," *Review of Quantitative Finance and Accounting* 1, 5–26.

Connor, Gregory, and Robert A. Korajczyk, 1993, "A Test for the Number of Factors in an Approximate Factor Model," *Journal of Finance* 48, 1263–1291.

Connor, Gregory, and Robert A. Korajczyk, 2009, "Factor Models of Asset Returns," Forthcoming, in *Encyclopedia of Quantitative Finance*, edited by Rama Cont. Chicester: Wiley, http://ssrn.com/abstract=1024709

Daniel, Kent, and Sheridan Titman, 1997, "Evidence on the Characteristics of Cross Sectional Variation in Stock Returns," *Journal of Finance* 52, 1–33.

Dhrymes, Phoebus J., Irwin Friend, and N. Bulent Gultekin, 1984, "A Critical Reexamination of the Empirical Evidence on the Arbitrage Pricing Theory," *Journal of Finance* 39, 323–346.

Elton, Edwin J., and Martin J. Gruber, 1973, "Estimating the Dependence Structure of Share Prices – Implications for Portfolio Selection," *Journal of Finance* 28, 1203–1232.

Elton, Edwin J., Martin J. Gruber, and Manfred W. Padberg, 1979, "Simple Criteria for Optimal Portfolio Selection: The Multi-Index Case," in *Portfolio Theory, 25 Years After: Essays in Honor of Harry Markowitz,* edited by E. J. Elton and M. J. Gruber. Amsterdam: North-Holland.

Epstein, Larry G., and Stanley E. Zin, 1989, "Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework," *Econometrica* 57, 937–969.

Epstein, Larry G., and Stanley E. Zin, 1991, "Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: An Empirical Analysis," *Journal of Political Economy* 99, 263–286.

Fama, Eugene F., and Kenneth R. French, 1992, "The Cross-Section of Expected Stock Returns," *Journal of Finance* 47, 427–465.

Fama, Eugene F., and Kenneth R. French, 1993, "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics* 33, 3–56.

Fama, Eugene F., and Kenneth R. French, 1996, "Multifactor Explanations of Asset Pricing Anomalies," *Journal of Finance* 51, 55–84.

Farrar, Donald E., 1962, *The Investment Decision Under Uncertainty*. Englewood Cliffs, NJ: Prentice-Hall.

Goyal, Amit, and Pedro Santa-Clara, 2003, "Idiosyncratic risk matters!," *Journal of Finance* 58, 975–1007.

Hansen, Lars Peter, and Kenneth J. Singleton, 1982, "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica* 50, 1269–1286.

Heaton, John, and Deborah J. Lucas, 1996, "Evaluating the Effects of Incomplete Markets on Risk Sharing and Asset Pricing," *Journal of Political Economy* 104, 443–487.

Heaton, John, and Deborah J. Lucas, 1999, "Stock Prices and Fundamentals," *NBER Macroeconomics Annual* 14, 213–242.

Heston, Steven L., and K. Geert Rouwenhorst, 1994, "Does Industrial Structure Explain the Benefits of International Diversification?" *Journal of Financial Economics* 36, 3–27.

Hopkins, Peter J. B., and C. Hayes Miller, 2001, *Country, Sector and Company Factors in Global Equity Models.* Charlottesvile, VA: The Research Foundation of AIMR and the Blackwell Series in Finance.

Huang, Chi-fu, and Robert H. Litzenberger, 1988, *Foundations for Financial Economics.* New York: Elsevier Science.

Jagannathan, Ravi, and Zhenyu Wang, 1996, "The Conditional CAPM and the Cross-Section of Expected Returns," *Journal of Finance* 51, 3–53.

Jegadeesh, Narasimhan, and Sheridan Titman, 1993, "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," *Journal of Finance* 48, 65–91.

Jegadeesh, Narasimhan, and Sheridan Titman, 2001, "Profitability of Momentum Strategies: An Evaluation of Alternative Explanations," *Journal of Finance* 56, 699–718.

King, Benjamin F., 1966, "Market and Industry Factors in Stock Price Behavior," *Journal of Business* 39, 139–190.

Kocherlakota, Narayana R., 1996, "The Equity Premium: It's Still a Puzzle," *Journal of Economic Literature* 34, 42–71.

Korajczyk, Robert A., and Ronnie Sadka, 2008, "Pricing the Commonality Across Alternative Measures of Liquidity," *Journal of Financial Economics* 87, 45–72.

Lintner, John, 1965, "Security Prices, Risk, and Maximal Gains from Diversification," *Journal of Finance* 20, 587–615.

Lucas, Robert E., Jr., 1978, "Asset Prices in an Exchange Economy," *Econometrica* 46, 1429–1446.

Markowitz, Harry M., 1952, "Portfolio Selection," *Journal of Finance* 7, 77–91.

Markowitz, Harry M., 1959, *Portfolio Selection: Efficient Diversification of Investments.* New York: Wiley.

Markowitz, Harry M., 1984, "The Two-Beta Trap," *Journal of Portfolio Management* 11, 12–19.

Markowitz, Harry M., 1987, *Mean–Variance Analysis in Portfolio Choice and Capital Markets.* Oxford: Basil Blackwell.

Mehra, Rajnish, and Edward C. Prescott, 1985, "The Equity Premium: A Puzzle," *Journal of Monetary Economics* 15, 145–161.

Merton, Robert C., 1973, "An Intertermporal Capital Asset Pricing Model," *Econometrica* 41, 867–887.

Mossin, Jan, 1966, "Equilibrium in a Capital Asset Market," *Econometrica* 34, 768–783.

Muirhead, Robb J., 1982, *Aspects of Multivariate Statistical Theory*. New York: Wiley.

Pástor, Ľuboš, and Robert F. Stambaugh, 2003, "Liquidity Risk and Expected Stock Returns," *Journal of Political Economy* 111, 642–685.

Roll, Richard, and Stephen A. Ross, 1980, "An Empirical Investigation of the Arbitrage Pricing Theory," *Journal of Finance* 35, 1073–1103.

Rosenberg, Barr, 1974, "Extra-Market Components of Covariance in Security Returns," *Journal of Financial and Quantitative Analysis* 9, 263–274.

Ross, Stephen A., 1976, "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory* 13, 341–360.

Ross, Stephen A., 1978a, "A Simple Approach to the Valuation of Risky Streams," *Journal of Business* 51, 453–475.

Ross, Stephen A., 1978b, "Mutual Fund Separation in Financial Theory – The Separating Distributions," *Journal of Economic Theory* 17, 254–286.

Shanken, Jay, 1987, "Nonsynchronous data and the covariance-factor structure of returns," *Journal of Finance* 42, 221–231.

Shanken, Jay, 1990, "Intertemporal Asset Pricing: An Empirical Investigation," *Journal of Econometrics* 45, 99–120.

Shanken, Jay, and Mark I. Weinstein, 2006, “Economic Forces and the Stock Market Revisited,” *Journal of Empirical Finance* 13, 129–144.

Sharpe, William F., 1963, “A Simplified Model for Portfolio Analysis,” *Management Science* 9, 277–293.

Sharpe, William F., 1964, “Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk,” *Journal of Finance* 19, 425–442.

Stock, James H., and Mark W. Watson, 2002, “Macroeconomic Forecasting Using Diffusion Indexes,” *Journal of Business and Economic Statistics* 20, 147–162.

Stroyny, Alvin L., 2005, “Estimating a Combined Linear Factor Model,” in *Linear Factor Models in Finance,* edited by J. Knight and S. Satchell. Oxford: Elsevier.

Tobin, J., 1958, “Liquidity Preference as Behavior Towards Risk,” *Review of Economic Studies* 25, 65–86.

Treynor, Jack L., 1961, “Toward a Theory of Market Value of Risky Assets,” Unpublished Working Paper.

Treynor, Jack L., 1999, “Towards a Theory of Market Value of Risky Assets,” in *Asset Pricing and Portfolio Performance: Models, Strategy, and Performance Metrics*, edited by Robert A. Korajczyk. London: Risk Publications.

Vassalou, Maria, 2003, “News Related to Future GDP Growth as a Risk Factor in Equity Returns,” *Journal of Financial Economics* 68, 47–73.

Weil, Philippe, 1989, “The Equity Premium Puzzle and the Risk-free Rate Puzzle,” *Journal of Monetary Economics* 24, 401–421.

Chapter 14
Applications of Markowitz Portfolio Theory To Pension Fund Design

Edwin J. Elton, Martin J. Gruber, and Christopher R. Blake

Harry Markowitz (1952) published an article that revolutionized the way the world thought about investments, the investment process, and measuring and predicting portfolio performance. The impact of Harry's 1952 article and his research that followed have been extraordinary. There is no doubt that this work is one of a handful of major breakthroughs that have changed not only the way we think about the world of finance, but also the way the world of finance functions.

The three of us have spent our entire careers teaching and working with the concepts that Harry set forth. When asked to do an essay for this volume, we were faced with an almost insurmountable problem: so many choices. We finally decided to review some of our works on pension fund management, because saving for retirement is a critical issue around the world and because the insight the profession has into the problem stems from Harry's work.

We start this essay with a brief introduction to 401(k) pension plans.

14.1 Introduction to 401(k) Pension Plans

Private pension funds in the United States have assets under management of over 14.3 trillion dollars. These retirement assets are held primarily in defined benefit plans, defined contribution plans, and individual retirement accounts. Defined

E.J. Elton (✉)
Stern School of Business, New York University, Department of Finance, Kaufman Management Center, 44 West 4th Street, Tisch 402, New York, NY 10012, USA
e-mail: eelton@stern.nyu.edu

M.J. Gruber
Stern School of Business, New York University, Department of Finance, Kaufman Management Center, 44 West 4th Street, KMC 9-88, New York, NY 10012, USA
e-mail: mgruber@stern.nyu.edu

C.R. Blake
Graduate School of Business Administration, Fordham University, 113 West 60th Street, New York, NY 10023, USA
e-mail: cblake@fordham.edu

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_14,
© Springer Science+Business Media, LLC 2010

contribution plans, and in particular 401(k) plans, are the fastest-growing type of retirement accounts in the United States.

These plans have certain key elements. First, the employer delineates a set of investment vehicles (e.g., mutual funds) among which an employee can allocate contributions. Employee contributions are made from before-tax income; and returns and contributions are not taxed until withdrawn. Usually the employer makes a contribution, the size of which is related to the employee's contribution.

The importance of these plans can be seen by the fact that more than one-third of American workers are covered by 401(k) plans. Furthermore, more than 60% of these workers have no other financial assets other than a bank account. In addition, for the largest group of 401(k) participants, those earning between $20,000 and $70,000 annually, the participants have less than one month's income invested outside of the plan. It is obvious that the value of 401(k) plans is of great importance to individuals and to society, for this value plays a key role in the adequacy of savings for retirement.

The value of any 401(k) pension plan to any participant is determined by two decisions: the set of investment choices offered to the participant in a plan, and how the participant allocates assets among these choices. These are two very different decisions. The first decision is made by the plan administrator; the second is made by the participant in the plan. There is a large amount of research on the participant's investment behavior, given the choices available to the participant.[1] However, none of this research examines the impact of the full set of investment choices offered in a plan on the ability of plan participants to construct desirable portfolios.

Examining the choices offered to the participants in any plan is important, because an investor faced with an inappropriate set of choices cannot construct an efficient portfolio from among these choices no matter what weights he places on the various offerings. There are two ways that the choices offered can be inappropriate: offering an insufficient number and type of choices to allow the construction of desirable portfolios, and offering poor-performing investment choices of any given type.

For that reason, we will divide the remainder of this essay into two sections. In the first section, we examine whether plan administrators offer their participants sufficient choices and the right mix of choices in the plan. In the second, we examine whether given the type of choices offered, administrators select good or bad investments for each type of choice selected.

14.2 The Adequacy of Investment Choices

Any 401(k) plan has a set of investment choices offered to the participants. We will examine the efficient frontier of these choices to find the optimum allocation decisions on the part of participants. We then ask the question whether adding a

[1] Examples of this extensive literature are Benartzi and Thaler (2001); Liang and Weisbenner (2002); Huberman and Sengmuller (2004); Agnew et al. (2003); Madrian and Shea (2001). This chapter is based on Elton et al. (2006).

reasonable set of investment alternatives to the set offered by plan administrators shifts the efficient frontier by a significant amount. We know that adding alternatives at random will shift the frontier. The question remains whether adding logical alternatives will shift it by an amount which did not arise by chance. To examine this, first we must construct a sample, second examine logical alternative choices for inclusion in a 401(k) plan, third examine a suitable methodology for determining whether a plan can be improved at a level, which is statistically significant, and finally present the results.

14.2.1 Sample

Our sample for this part of the study consists of data for the year 2001 on 417 different 401(k) plans that held only publicly available mutual funds with at least 5 years of monthly return data. The median number of offerings by any plan is eight with 12% offering four or fewer choices and 11% offering more than 12 choices.

14.2.2 Alternative Investment Choices

In this section, we examine the adequacy of the investment choices offered by 401(k) plans. In order to determine if 401(k) plans offer their participants appropriate investment choices, we need to hypothesize an adequate set of alternative investment choices. We construct the alternatives by drawing on the literature of financial economics, which discusses indexes that are necessary and sufficient to capture the relevant return characteristics for a range of investments.[2]

The indexes employed will now be described in greater detail. For common stocks, we classify by value versus growth and by size as advocated by Fama and French (1995). We classified the size into two groups: small-mid-cap and large-cap. Each of these two groups was then further divided into value and growth. All four indexes were taken from Wilshire. We chose Wilshire indexes because there exist tradeable funds that attempt to match each of the indexes. For bonds, we combined a general bond index, including governments and corporates, and a mortgage-backed index. We also employed a high-yield index. This division is supported by Blake et al. (1993), who found this division sufficient to capture differences in return across bond funds. For the combined bond index, we used the Lehman U.S. Government/Credit Index and the Lehman Fixed-Rate Mortgage-Backed Securities Index. We used the Credit Suisse First Boston High-Yield Index for the high-yield

[2] We compared the commonly used industrial classifications as indexes with the indexes from financial economics and found that the indexes from the literature were superior in explaining returns.

bond index.[3] We also included the Salomon Non-U.S. Dollar World Government Bond Index for international bonds and MSCI EAFE Index for international stocks.

Since, in this study, returns on all mutual funds are computed after expenses, we deducted expenses from each of our indexes. For each of our indexes, we used the expense charge of the index fund (including exchange-traded funds) that most closely matched the index. If there were multiple index funds matching the index, we would have used the expense charge of the lowest cost fund.

We now examine whether the choices given to investors allow construction of an efficient frontier similar to that obtained by the eight indexes. To do this we use intersection tests.

14.2.3 Methodology

The purpose of the intersection test is to examine whether, given a riskless rate, a particular set of assets is sufficient to generate the efficient frontier or whether including (long or possibly short) members of a second set of assets would improve the efficient frontier at a statistically significant level. In other words, given the best portfolio (linear combination) of the assets held by a plan, can adding one or more of the eight indexes to this portfolio statistically significantly improve the return for the portfolio at a given level of risk, or do the original plan offerings span the eight indexes?

As DeRoon et al. (2001) have shown, intersection is a test of the impact of restricting the intercept (a) in the following time-series model:

$$R_{it} - R_f = \alpha_i + \sum_{k=1}^{K} \beta_{ik}(R_{it} - R_f) + \varepsilon_{it} \tag{14.1}$$

Where

R_{it}= the return on one of the eight indexes, described earlier, in month t ($i = 1, \ldots, 8$),

R_f = the risk-free rate,

R_{kt} = the return on fund k in month t,

ε_{it} = the error term for index i in month t,

β_{ik} = the sensitivity of index i to plan fund k.

Investors can only hold 401(k) plan assets and additional mutual funds long. When short sales are not allowed, the right-hand side of (14.1) includes returns only on those funds in a plan that is held long in the optimal portfolio of plan funds.[4]

[3] Originally, the indexes for the mortgage-backed securities and small- and mid-cap stocks were included separately, but empirical tests showed there were no improvements in explaining returns by including them separately.

[4] These assets can be easily identified by solving a quadratic programming problem for each plan.

Intersection occurs if, for all of the eight indexes jointly, the a_1 is not statistically significantly positive, i.e., the restrictions are

$$\alpha_i \leq 0 \forall i$$

The logic behind the test can be easily understood.[5] The intercept is the additional risk-adjusted return that is available on each of the eight indexes that is not available on a linear combination of the plan assets. If short sales are forbidden, then only the addition of an asset with a positive alpha can improve the efficient frontier, by offering a higher return on an optimal portfolio for any given level of risk. Since most assets would have a positive or negative alpha, the test examines whether the shift in the efficient frontier associated with positive alphas is statistically significant. Since adding assets to any portfolio has a probability of improving efficiency, it is important to employ a test of statistical significance. The test we employ is the intersection test, given a riskless rate and short sales not allowed, shown in DeRoon et al. (2001).

14.2.4 Results

The results of the intersection tests are shown in Table 14.1. Recall that the intersection tests would not reject spanning if adding any of the eight indexes to the optimal portfolio of funds offered by any plan does not improve the efficient frontier at a statistically significant level. Plans, where spanning is not rejected, offer participants a sufficient set of choices. The majority of plans holding four or fewer funds do not offer a set of funds that span the eight indexes. For these plans, there are more indexes than fund offerings. However, it is possible that a small set of funds spans the larger set of indexes, either because some of the indexes are not desirable investments or because some of the funds are combinations of two or more of the indexes. However, this does not happen for many plans offering a small set of investment choices. For plans holding seven or more funds, we find that about 60% of the plans offer investment choices that span the relevant space investors are interested in. Of course, the glass is also about half empty and in that 40% of the plans may leave investors unsatisfied. Finally, it is not until plans offer 14 or more investment choices (4.3% of all plans) that virtually all plans offer investment choices that span the space investors should be interested in. Of the 417 plans, only 53% span the space obtainable from the eight indexes. While some 401(k) plans offer participants a rich enough selection of investment choices to satisfy their needs, clearly a number of 401(k) plans do not do so.[6]

---

[5] For the theory and application of spanning tests, see, for example, Bekaert and Uris (1996); Chen and Knez (1996); Hansen et al (1995); Huberman and Kandel (1987) and Glen and Jorion (1993). For foundations see Markowitz (1956 and 1959).

[6] As a further check on plans spanning, we considered whether plans spanned the space of the simplest set of choice we should think of: a broad stock market index (the Wilshire 5000 index), a broad market bond index (a combination of the Lehman U.S. Government/Credit index and the

Table 14.1 Sufficiency of plan investment choices in spanning eight indexes (short sales not allowed)

Number of investment choices in plan	Total number of plans	Number of plans that span (offer sufficient choices)
1	10	3
2	18	4
3	37	18
4	57	28
5	53	33
6	58	24
7	44	26
8	39	22
9	45	25
10	14	8
11	11	8
12	11	5
13	2	1
14	4	3
15	7	7
16 or more	7	6
Total	417	221

This table shows the total number of sample plans based on number of investment choices offered (excluding company stock, money market funds, GICs and stable value funds), along with the number of plans within each total that span eight indexes. A plan that spans offers a sufficient set of choices, so that the plans' optimum tangent portfolio is not improved by including one or more of the eight indexes. The eight indexes consist of four domestic equity indexes, one international equity index, two domestic fixed-income indexes, and one international fixed-income index. The sample period covers the 5 years from January 1997 through December 2001. Monthly return data for the investment choices in the 401(k) plans were obtained from the CRSP databases

Before leaving this section, it is worthwhile examining the loss in return to 401(k) plan holders due to plans not spanning the relevant space. To measure this, we employ the Sharpe ratio. Recall that the Sharpe ratio is equal to the return on a particular portfolio minus the riskless rate, all divided by the standard deviation of the portfolio. The Sharpe ratio is the standard measure of the efficiency of a portfolio in mean standard deviation space. Two portfolios with the same ratio are equally desirable. Thus for each plan in our sample we can ask the question: given the optimum portfolio for a plan that does not span, how much higher would the return have to be to give the same Sharpe ratio as the optimal portfolio composed of the eight indexes?

Lehman Mortgage-Backed Securities index), and an international index (the MSCI EAFE Index). We adjusted the returns of the three indexes to reflect normal management fees (just as we did for the eight indexes). With this limited set of three indexes, more plans offered choices that spanned the indexes' space (236 of the 417 plans).

In order for the 196 plans in our sample, which did not span to have the same Sharpe ratio as the optimum portfolio comprised of the eight indexes, the average return on the plans' optimum portfolios, holding risk constant, would have to increase by 0.178% per month. This means that monthly returns would have to increase from 0.866% to 1.044% per month. For an investor who remains in a plan for 20 years, this results in ending wealth that is 53% higher than he or she would receive in a plan with insufficient options.[7] Thus investors in 401(k) plans are sacrificing significant returns because plan administrators are offering an incomplete set of investment alternatives.[8]

14.3 Administrator Selection of Individual Funds

We have seen that pension plan administrators on an average offer participants a few types of mutual funds in their plans, too. The second question of interest is given the type of funds offered; do administrators pick good funds relative to the population of choices that are available?

14.3.1 Sample

Our sample for this part of the study consisted of detailed information on 289 plan years from 11-K filings of holdings and aggregate participant allocation for the years 1994 through 1999 in mutual funds investments. Data are not available after 1999 due to change in SEC requirements.

The sample contained data on 43 individual pension plans, most of which had 7 years of data and all of which held only mutual funds or cash accounts. This sample differs from that used in the first part of this study in which while it contained fewer plans, it contained much more data over time for each plan. In addition, this new data contains information on participant asset allocation.

14.3.2 Performance

We examine three aspects of performance in this section of the paper, namely whether the given type of funds offered, 401(k) plans offer superior-performing mutual funds, whether they improve performance when they change offerings, and

[7] Ending wealth would be $(1.00866)^{240}$ versus $(1.01044)^{240}$ or \$7.92 versus \$12.09 for every dollar invested.

[8] These differences are much larger than any possible differences due to expense ratios between index funds and active portfolios. See **?** (**?**) for estimates of expense ratios and Elton et al. (2004a) for index fund expenses.

whether some plan administrators are better than others. The first issue, we address, is the performance of the funds selected by the plan administrators.

14.3.2.1 Fund Performance

How well do the funds selected by plan administrators perform compared to what they could have selected? To analyze this, we first compute alphas on the funds that were offered to participants. Alpha is a measure of the return the participant earns above or below what she would have earned if she had simply held a passive portfolio of indexes with the same risk.

Alpha is estimated from the followed regression:

$$R_{it} - R_{ft} = \alpha_i + \sum_{k=1}^{K} \beta_{ik} I_{kt} + \varepsilon_{it}$$

Where

1. R_{it} is the total return on mutual fund i in period t.
2. β_{ik} is the sensitivity of fund i to index k.
3. R_{ft} is the riskless rate in period t.
4. ε_{it} is the error term for fund i in period t.
5. I_{kt} in period t is either the return on index k, if it is the difference between two return series, or the return on index k above the riskless rate if it is a single return series.

The overwhelming evidence is that alpha on an average is negative for mutual funds.[9] Thus a negative alpha on an average offering 401(k) plans is not inconsistent with plan administrators picking better-performing actively managed mutual funds. Therefore, to evaluate the alpha on the funds in a plan we will compute for each fund in the plan a "differential alpha," which is the difference between alpha on each fund in a plan and the average alpha on a random sample of similar funds. We define similar funds as funds in the same ICDI investment objective category of similar size. Adjusting for ICDI category controls year-to-year variation in the performance of any category not captured by our performance model. Adjusting for the size of the mutual fund controls for any difference in fund performance is due to the size of funds.[10] The differential alpha is the primary metric that we will use to judge the ability of pension fund administrators to select mutual funds with superior performance.

---

[9] See Blake et al. (1993); Grinblatt and Titman (1989); Jensen (1968); Sharpe (1966) and Wermers (2000).

[10] The adjustment for size and category is performed by taking a random sample of 100 funds (or the maximum available if there are less than 100) in the same ICDI category, dividing this sample into deciles based on size, and comparing the fund's alpha with the average alpha from the size decile (in its ICDI category) into which the fund falls.

For purposes of specifying the appropriate indexes for defining alpha, mutual funds will be divided into three types: stock funds, bond funds, and international funds. For stock funds we will use a five-index model, the S&P 500 index, both Fama/French small-minus-big (SMB) index and high-minus-low (HML) indexes, the Lehman Government/Credit index and MSCI EAFE index. Two indexes require some comments. First, the bond index is needed because the stock category includes many funds that are combinations of bonds and stocks, such as balanced and income funds, and because funds in the common stock categories, such as aggressive growth or long-term growth, often hold part of their portfolio in long-term bonds. Failure to include a bond index imputes to alpha any return on long bonds different from the riskless rate. The other nonstandard index is the international index. During the period of this study, many stock funds included international stocks in their portfolio, usually in the form of ADRs. Again, failure includes an international index that would cause fund alphas to be a function of their international holdings.

For bond funds, we will use a four-index model: the Lehman Government/Credit Index, the Lehman Fixed-Rate Mortgage-Backed Securities Index, the Credit Suisse/First Boston High-Yield Index, and the Citigroup Non-Dollar World Government Bond Index. The first three indexes are supported by the work of Blake et al. (1993). The addition of the international index is needed to capture the tendency of some bond funds to include international bonds in the portfolio over this period.

Finally, for international funds we will use a five-index model consisting of the S&P 500 index (since international funds often hold the U.S. securities), three MSCI indexes (Europe, Pacific and Emerging Markets), and the Citigroup Non-Dollar World Government Bond Index.

We computed monthly alphas for 1 and 3 years following the date of each 11-K report. Using subsequent return data to evaluate performance eliminates the bias that results from using data prior to the 401(k) report since, as shown later, plans add new funds after period of superior past performance. Thus evaluation using the period prior to the report would pick up returns not available to participants (if the fund was newly added) and would bias 401(k) plans' relative alphas upward. Three-year alphas are estimated over the 36 months following the fiscal year end. For 1-year alphas, we take the fund's 3-year alpha and added to it the average monthly residual over the 12 months following the date of the 11-K report.

To compute the alpha on a plan from fund alphas, we will use two different weighting schemes. First, we examine the average performance of each plan administrator. This involves an equal weighting of each fund in the plan. Second, we weight the funds unequally using the actual weights chosen by participants as shown in the 11-K statement. This is an appropriate criterion for plan administrators, if they forecast correctly on the types of funds participants will invest in and allocate more of their efforts to selecting individual funds of those types. These weights are determined by dividing the amount invested in any fund by plan participants by the aggregate amount invested in all funds in the plan. Table 14.2 shows the results for 3-year alphas, and Table 14.3 shows the results for 1-year alphas.

Tables 14.2 and 14.3 present data on the selection ability of pension fund administrators. First, note that the alphas for the average plan are negative, and consistent

Table 14.2 3-Year 401(k) plan performance results (monthly alphas)

		Using equal weights on funds in plan		Using participant weights on funds in plan	
Plan	No. of years	3-year Alpha	3-year Differential alpha	3-year Alpha	3-year Differential alpha
1	8	0.075%	0.169%	0.117%	0.202%
2	7	0.025%	0.074%	0.032%	0.121%
3	7	0.039%	0.099%	0.028%	0.096%
4	8	−0.091%	−0.036%	−0.208%	−0.126%
5	6	−0.035%	0.097%	−0.051%	0.093%
6	7	−0.076%	−0.002%	−0.050%	0.023%
7	5	−0.070%	0.012%	−0.080%	0.014%
8	7	0.092%	0.115%	0.126%	0.142%
9	8	−0.064%	−0.011%	−0.064%	−0.029%
10	7	0.057%	0.111%	0.062%	0.094%
11	4	−0.070%	0.053%	−0.079%	0.033%
12	6	0.021%	0.075%	0.002%	0.055%
13	7	−0.058%	−0.014%	−0.124%	−0.083%
14	8	−0.036%	−0.014%	−0.061%	−0.004%
15	7	0.020%	0.088%	−0.009%	0.083%
16	7	−0.071%	−0.018%	−0.056%	0.016%
17	7	−0.055%	0.019%	−0.070%	0.008%
18	8	0.059%	0.109%	0.005%	0.089%
19	8	−0.072%	−0.013%	−0.084%	−0.041%
20	4	−0.049%	0.074%	−0.052%	0.119%
21	5	−0.020%	0.106%	−0.019%	0.113%
22	5	0.043%	0.190%	0.107%	0.271%
23	4	−0.481%	−0.333%	−0.399%	−0.254%
24	5	−0.002%	0.165%	−0.012%	0.163%
25	7	0.120%	0.152%	0.129%	0.166%
26	7	0.034%	0.088%	0.006%	0.081%
27	7	−0.001%	−0.030%	−0.021%	0.017%
28	8	0.111%	0.161%	−0.012%	0.071%
29	4	0.002%	0.103%	−0.032%	0.097%
30	6	−0.164%	−0.038%	−0.164%	−0.038%
31	4	0.170%	0.029%	0.170%	0.029%
32	8	−0.043%	0.006%	−0.045%	0.000%
33	8	0.078%	0.141%	0.087%	0.137%
34	8	0.035%	0.081%	−0.031%	0.019%
35	7	0.061%	0.099%	0.077%	0.151%
36	7	−0.128%	−0.017%	−0.115%	−0.010%
37	8	−0.054%	0.006%	−0.047%	0.003%
38	9	−0.055%	0.056%	−0.035%	0.072%
39	8	0.053%	0.119%	−0.020%	0.065%
40	8	−0.003%	0.051%	−0.056%	0.030%
41	8	−0.068%	−0.018%	−0.148%	−0.099%

(continued)

Table 14.2 (continued)

		Using equal weights on funds in plan		Using participant weights on funds in plan	
Plan	No. of years	3-year Alpha	3-year Differential alpha	3-year Alpha	3-year Differential alpha
42	7	−0.001%	0.027%	−0.074%	−0.003%
43	5	−0.464%	−0.295%	−0.570%	−0.396%
Average	6.721	−0.026%	0.043%	−0.043%	0.037%
Two-tail p-value		0.161	0.009	0.034	0.040

This table shows plan performance results using weighted averages of monthly alphas calculated over a 3-year regression period. One weighted average uses equal weights in all choices in a given plan's fiscal year; the other weighted average uses participants actual weights at the end of a given fiscal year. For each fiscal year in a plan and for each fund in the plan for that year, a fund's 3-year alpha is the intercept from a 36-month multi-index regression of the fund's monthly excess returns on a set of indices determined by the funds type (stock, bond or international), starting in the month following the fiscal year end. (See text for regression model details.) Differential alpha is the fund's alpha minus the average alpha of funds in the matched size decile from a random sample of funds with the same investment objective as plan's fund.

with results found for mutual funds in other studies. The magnitude of these negative alphas is larger than normal expenses for low-cost index funds, suggesting that performance would be improved if passive funds had been substituted for the active funds which were selected. However, examining differential alpha, it is clear from these tables that pension fund administrators have selected mutual funds for their participants and that are better than a set of randomly chosen mutual funds of the same type, size and risk. Examining Table 14.3 shows that using equal weights for the funds in each plan, over a 3-year period plan administrators had an average differential alpha over funds of the same type, size and risk of 4.3 basis points per month, or approximately 52 basis points per year. As shown in Table 14.3, over a 1-year period the equal-weighted differential alpha is 3.5 basis points per month, or approximately 42 basis points per year. Both of these differential alphas are statistically significantly different from zero at the 5% level. In addition, the 3-year differential alpha is positive for 30 out of 43 plans, and the 1-year differential alpha is positive for 29 out of 43 plans.

The last two columns of Table 14.2 weight the performance of any fund in a plan by the fraction of assets participants place in that fund. The weighted performance is a result of both the funds offered to plan participants by plan administrators and the choices among those funds made by participants in the plan. Using participants' actual weights is a relevant alternative to equal weighting. Perhaps management is selecting funds based on forecasts of how participants will weight the funds selected, and thus devoting more effort to selecting funds expected to be heavily invested in. The numeric results change very little by using participant weights rather than equal

Table 14.3 1-Year 401(k) plan performance results (monthly alphas)

		Using equal weights on funds in plan		Using participant weights on funds in plan	
Plan	No. of years	1-year Alpha	1-year Differential alpha	1-year Alpha	1-year Differential alpha
1	8	−0.052%	0.104%	−0.047%	0.119%
2	7	−0.051%	0.081%	−0.044%	0.123%
3	7	−0.044%	0.100%	−0.030%	0.123%
4	8	−0.171%	−0.074%	−0.345%	−0.202%
5	6	−0.062%	0.066%	−0.090%	0.055%
6	7	−0.080%	0.055%	−0.056%	0.099%
7	5	−0.204%	0.012%	−0.111%	0.087%
8	7	0.017%	0.101%	0.006%	0.102%
9	8	−0.133%	−0.033%	−0.177%	−0.080%
10	7	0.003%	0.104%	0.032%	0.111%
11	4	−0.163%	−0.013%	−0.174%	−0.035%
12	6	−0.058%	0.101%	−0.058%	0.105%
13	7	−0.140%	−0.020%	−0.258%	−0.126%
14	8	−0.089%	−0.011%	−0.111%	0.008%
15	7	−0.022%	0.101%	−0.035%	0.121%
16	7	−0.143%	−0.008%	−0.083%	0.089%
17	7	−0.130%	0.020%	−0.111%	0.066%
18	8	−0.021%	0.091%	−0.029%	0.119%
19	8	−0.093%	−0.014%	−0.064%	0.034%
20	4	−0.054%	0.036%	−0.041%	0.084%
21	5	−0.025%	0.062%	−0.024%	0.069%
22	5	−0.113%	0.199%	−0.109%	0.223%
23	4	−0.453%	−0.420%	−0.300%	−0.270%
24	5	−0.014%	0.135%	−0.014%	0.143%
25	7	0.049%	0.141%	0.050%	0.161%
26	7	−0.020%	0.127%	−0.039%	0.130%
27	7	0.001%	0.035%	−0.019%	0.098%
28	8	0.037%	0.166%	−0.040%	0.130%
29	4	−0.021%	0.076%	−0.002%	0.111%
30	6	−0.138%	−0.059%	−0.138%	−0.059%
31	4	0.228%	0.106%	0.228%	0.106%
32	8	−0.122%	−0.023%	−0.165%	−0.036%
33	8	0.004%	0.104%	−0.010%	0.105%
34	8	0.012%	0.079%	−0.081%	−0.005%
35	7	−0.019%	0.079%	−0.023%	0.132%
36	7	−0.155%	0.006%	−0.173%	0.020%
37	8	−0.106%	−0.003%	−0.092%	0.018%
38	9	−0.102%	0.005%	−0.058%	0.041%
39	8	0.019%	0.146%	−0.036%	0.135%
40	8	−0.014%	0.083%	−0.081%	0.057%
41	8	−0.169%	−0.061%	−0.310%	−0.190%

(continued)

Table 14.3 (continued)

		Using equal weights on funds in plan		Using participant weights on funds in plan	
		1-year	1-year	1-year	1-year
	No. of		Differential		Differential
Plan	years	Alpha	alpha	Alpha	alpha
42	7	−0.133%	−0.017%	−0.136%	0.025%
43	5	−0.492%	−0.255%	−0.576%	−0.365%
Average	6.721	−0.080%	0.035%	−0.093%	0.041%
Two-tail p-value		0.000	0.038	0.000	0.029

This table shows plan performance results using weighted averages of monthly alphas over a 1-year period. To obtain 1-year monthly alphas for each fiscal year in a plan and for each fund in the plan for that year, the intercept from a 36-month multi-index regression of the fund's monthly excess returns on a set of indices determined by the fund's type (stock, bond or international), starting in the month following the fiscal year end, is added to the average monthly residual over the first 12 months of the regression period. One weighted average uses equal weights in all choices in a given plan's fiscal year; the other weighted average uses participant's actual weights at the end of a given fiscal year. Differential alpha is the fund's alpha minus the average alpha of funds in the matched size decile from a random sample of funds with the same investment objective as plan's fund.

weights. With participant weighting, the average 3-year differential alpha drops from 4.3 basis points to 3.7 basis points and is positive for 32 of 43 plans, while the 1-year differential alpha increases from 3.5 to 4.1 basis points and is positive for 33 plans.

It is clear from these results that plan administrators possess some skill in selecting funds for their plans.[11] While they exhibit some skill in choosing among actively managed funds, they would produce better results for plan participants by choosing index funds rather than actively managed funds. Furthermore, since participant investment allocation in aggregate seems to neither add nor to subtract from the performance attributed to plan administrators, the key element in how well the participants in the aggregate fare appears to be the choices that the participants are offered. Participants do not increase their performance results compared to investing an equal amount in each mutual fund by their allocation choices.

14.3.2.2 Performance of Additions and Deletions

Pension fund administrators change the investment choices offered to participants with surprising frequency. This is particularly true with respect to the inclusion of new mutual funds. Across the 289 fund years in our sample, there were 215 mutual

[11] Part of the skill may be in selecting lower cost funds. Another possible explanation is that funds made available for 401(k) plans are lower-cost funds.

funds added to the offerings and 45 funds dropped. The pattern of deletions is interesting. First, very often, dropping a fund seems to be part of a more general change in direction for a plan rather than an isolated decision on the performance of a single fund. Only in six cases was a single fund dropped from a plan. In the other 39 incidents of dropped funds, two or more funds were dropped from a plan simultaneously. In fact, for 24 of the 45 funds that were dropped, the plan that held them dropped all the funds in the plan simultaneously. Second, in many cases dropping a fund did not seem to be motivated by a desire to pick a better manager of the same type, since 20 dropped funds were not replaced with funds in the same category and, for 15 of those 20 cases, the ICDI category of the dropped fund was completely eliminated from the plan.

Many of the additions seem to be motivated by a desire to add a new type of fund. Of the 215 funds added to plans over our sample period, 146, or over half, were selected from an ICDI investment objective category not held by the plan at the time of the addition.

What are the characteristics of funds that are added and deleted? Panel A of Table 14.4 presents data on funds that were added or dropped for the 1- and 3-year periods prior to that change. A fund is considered dropped at the beginning of the first year and it does not appear in the 11-K annual statement; it is considered added at the end of the first year in which it appears in the 11-K annual statement. For measuring performance prior to a fund being added or dropped, we ran regressions ending one year prior to the fiscal year date in which the fund appeared or disappeared. The regression period covered a minimum of 18 months up to a maximum of 36 months.

From Table 14.4 we see that funds that were added have non-negative differential alphas for both 1 year and 3 years prior to the change. In contrast, funds that were dropped have negative differential alphas both 1 year and 3 years before they were dropped. Except for the 1-year results for dropped funds, for both added and dropped funds the differential alphas are statistically significantly different from zero at the 5% level. Note that the difference in differential alpha between the added and dropped funds is statistically significant at the 1% level for the 1- and 3-year periods before the change. Also, the funds that were added have a differential alpha above those that were dropped of 23 basis points per month, or 276 basis points per year, for 3 years before the change and 19 basis points per month, or 228 basis points per year, for 1 year before the change. Table 14.4 also shows results for alpha and differential return. We see that, for both of these other measures, the sample of added funds has much better past performance than the sample of dropped funds. The decision to add or drop funds would seem to be made, at least in part, on the basis of past performance.

Since plans frequently add new types of funds, it is interesting to examine whether plan administrators are adding "hot" investment objectives (types). Anytime a fund was added to a plan, we calculated the alpha for the random sample from the ICDI category the fund was in for that year. This is compared to the average alpha over the total random sample (all types) for that year. Over the 3 years prior to the addition, the average difference in monthly alpha is 1.8 basis points

Table 14.4 Fund performance before and after being added or dropped (monthly data)

	Panel A: before being added or dropped		
	Differential		Differential
Added funds (193)	Return	Alpha	Alpha
1 year	0.251%	0.141%	0.115%
3 year	0.222%	0.190%	0.112%
Dropped funds (41)			
1 year	−0.024%	−0.137%	−0.077%
3 year	−0.028%	−0.124%	−0.119%
Added minus dropped funds			
1-year difference (p-value)	0.275% (0.000)	0.278% (0.000)	0.192% (0.001)
3-year difference (p-value)	0.250% (0.000)	0.314% (0.000)	0.231% (0.000)
	Panel B: after being added or dropped		
	Differential		Differential
Added funds (193)	Return	Alpha	Alpha
1 year	0.008%	−0.142%	−0.022%
3 year	0.144%	0.010%	0.037%
Dropped funds (41)			
1 year	0.190%	−0.062%	0.086%
3 year	0.073%	−0.037%	0.014%
Added minus dropped funds			
1-year difference (p-value)	−0.182%	−0.080%	−0.108%
3-year difference (p-value)	0.071% (0.285)	0.047% (0.558)	0.023% (0.739)

For each plan year in which a fund is added or dropped, "before" and "after" performance measures are calculated going backward from 1 year before the fund is dropped or forwarded from the month in which the fund is added.
Differential return is each fund's average monthly return minus the average monthly return of a random sample of funds with the same investment objective.
Alpha is the intercept from a multi-index regression of the fund's monthly excess returns on a set of indices determined by fund's type (stock, bond or international). (See text for regression model details.) The 1-year alpha is the fund's 3-year alpha plus the average monthly residual over the relevant period.
Differential alpha is each fund's alpha minus the average alpha of funds in the matched size decile from a random sample of funds with the same investment objective.

(significant at the 1% level). Plan administrators are adding an ICDI investment objective category that performed better in the past than the average category.

A logical question to ask is how well funds do after they are added or dropped. These results are presented in the lower panel of Table 14.4. Over 3-year periods and 1-year period (starting in the month after the end of the fiscal year in which the fund appears or disappears), for all three measures of performance (differential return, alpha and differential alpha), the performance of the added funds is no better than the performance of the dropped funds at any reasonable level of statistical significance, and none of the differential alphas are statistically significantly different from zero. Also note that the differential alpha of the added funds after they are added is worse than their performance before they are added, while the performance of the dropped funds is better after they were dropped compared to their performance before. These results are true for both the 1- and 3-year analyses, and the results are statistically significant at the 1% level.

For further evidence, we examined two specific sub-samples. First, we examined the 26 cases were a 401(k) plan deleted a fund and replaced it with a fund with an identical objective. In this sub-sample, the deleted funds outperform the funds they replace by an average of 20 basis points per month over the next 3 years, and by an average of three basis points per month over the 1 year.

Second, we examined the sub-sample of the nine plans that changed virtually all of their offerings by dropping funds from one fund family and adding funds from a new fund family.[12] Since this involved replacing an entire portfolio with another portfolio, an aggregate measure of performance is appropriate. To measure overall performance, we calculated future Sharpe ratios for each plan using both equal weights and participant weights. Table 14.5 shows results both for 1-year and 3-year Sharpe ratios, assuming equal weights. Under either set of weights, the past Sharpe ratios are higher for the portfolio of added funds than for the portfolio of dropped funds, while, after replacement, the future Sharpe ratios are higher for the portfolio of dropped funds than for the funds that replaced them.[13]

The overall pattern is clear. Plan administrators add funds that have done well in the past and drop funds that have performed poorly. However, when examining future performance the dropped funds seem to perform no worse than the added funds.[14]

Table 14.5 Sharpe ratios of 401(k) plans that changed all fund offerings in a given year (monthly data)

	Before change	After change
Added		
1 year	0.325	0.393
3 year	0.309	0.223
Dropped		
1 year	0.283	0.479
3 year	0.257	0.253

Average Sharpe ratios are based on equal amount invested in each fund in a plan for a given year.
Future 3-year Sharpe ratios are calculated over a period of 36months starting from the month following the fiscal year end (FYE) in which the fund was added or dropped; future 1-year Sharpe ratios are calculated over a 12-month period starting from the month following the FYE in which the fund was added or dropped. Past 3-year Sharpe ratios are calculated over a 36-month period ending in FYE month; past 1-year Sharpe ratios are calculated over a 12-month period ending in the FYE month.

[12] In two of the cases, a plan replaced all, but one fund; in all other cases, all funds were replaced.

[13] Table 5 shows results where each fund in a plan was equally weighted; the results were very similar when funds were weighted using participant allocations.

[14] Additional analysis reveals that participants invest less in added funds than in funds that continue to be offered and that the less added funds resemble funds offered in the past, the more reluctant participants are to invest in them.

14.3.2.3 Identification of Superior Plan Administrators

Having shown that, on an average, plan administrators select funds that perform better than average, holding constant the ICDI classification and size, we now examine whether past performance allows us to identify plans that have superior performance in the future.

Before doing so we examine whether, for the entire sample of plans, certain plan administrators show performance that is different from other plan administrators at a statistically significant level. An analysis of variance test was performed using the entire history for all funds. The results showed a probability less than 0.01 (p-value of 0.0013) that there was no difference in the ability of individual plan administrators.

We now turn to an examination of whether good performance of a plan in one period predicts good performance in a second. In order to have non-overlapping periods, we will confine the analysis to 1-year alphas. We will continue to use the performance metric we have described as most relevant differential alpha. The results are shown in Table 14.6. To construct Table 14.6 for each year from 1991 through 1998, the 401(k) plans in our sample are ranked and divided into four equal groups based on past 1-year differential alphas. For each year, the fraction of plans that stay in their quartile or migrate to each of the other quartiles is calculated. The overall results are shown in Table 14.6 along with the actual average differential alpha that occurred in the year after the ranking. If there was no persistence, the diagonal elements should all be equal to 0.25. Examining Table 14.6 shows that they are all above 0.25 and they average 0.3375. We can get a better idea of the persistence of this performance by examining the differential alpha that would have been earned in the subsequent year by holding plans in each quartile. The differential alpha from holding the bottom quartile of plans is −0.024 or approximately

Table 14.6 Predictability of future performance from past performance

Past and future performance quartiles

Past Performance Quartiles	Future performance quartiles 1 (lowest)	2	3	4 (highest)	Average future differential alpha
1 (lowest)	0.338	0.265	0.235	0.162	−0.024%
2	0.203	0.297	0.216	0.284	0.040%
3	0.162	0.203	0.419	0.216	0.063%
4 (highest)	0.254	0.254	0.197	0.296	0.061%

Plan performance for a given year is measured using the average 1-Year differential monthly alpha of the funds in the plan that year, where a fund's monthly differential alpha is the fund's 1-year monthly alpha minus the average 1-year monthly alpha of funds in the matched size decile from a random sample of funds with same investment objective as plan's fund.

The table shows both the probabilities of plans being in future performance groups, conditional on being in a past performance group, and the average future performance of plans in past performance groups.

All results are calculated over 287 plan years. (Two plan years in 1999 were dropped from sample of 289 plan years since only 2 plans were in 1999 sample.)

−0.29 basis points per year, while for the top quartile it is 0.061 basis points or approximately 0.74 basis points per year. The difference is statistically significant at the 0.01 level. Differences in future differential alphas between the top three quartiles are insignificant.[15]

While there is some predictability of future performance of plans from past performance, most of that comes from bad plan performance predicting bad plan performance.[16]

14.4 Conclusion

Harry Markowitz' pioneering work in portfolio analysis has changed the way we think about the investment process and has allowed us to gain insight into the economic efficiency of financial institutions.

One of the most important applications of Markowitz' work is in the area of pension design. Given the aging of the population in almost all countries around the world and the increased recognition that government-provided pension protection cannot meet the needs of future generations, the design of appropriate private pension plans becomes a critical issue. In this essay, we have selected one of the most important types of private pension plans, 401(k) plans, to investigate the efficiency with which these plans are administered. Recall that in a 401(k) plan the administrators select a set of investment possibilities and the plan participant is free to choose from among this restricted set of choices.

In this essay, we examine the reasonableness of the investment choices offered by 401(k) plans. This is an important subject. The payoff of a pension plan to any investor is the product of two different decisions: what the investor is offered, and what he or she chooses from what is being offered. While a lot of attention has been paid to participant choice, little attention has been paid to the relevancy of the choices offered to participants. If investors are given an inferior set of choices in their plan, the effectiveness of their choices is severely constrained. The efficiency

[15] Some of the persistence in plan performance could be due to mutual fund persistence. Most of the bad future performance and persistence comes from the bottom decile. As a final step, we examined whether better performance was associated with plan characteristics. If it were, we could identify plans that would perform well in the future. To examine this, we regressed the average alpha for a plan on the log of the average dollar size of the plan, the average number of choices every year, the average number of changes in choices every year, the average new cash flows divided by the size of the plan, dummy variables for the presence of money market funds and GIC's, and a dummy variable representing whether or not the company offering the plan was a high-tech or financial firm. None of the variables' coefficients were close to being statistically significant except for the high-tech/financial firm dummy, which was significant at the 10% level.

[16] Research has shown that bad-performing mutual funds in one period tend to be bad-performing funds in the next period. Recall, however, that we are examining predictability of plans that are portfolios of funds, which, on average, outperform random selection. For this group, it is not clear without empirical evidence that bad performance of plans predicts future bad performance of plans.

of the set of choices offered to participants has two parts: the adequacy of the type of and number of choices offered to plan participants and the ability of the plan administrator to select individual investment vehicles to represent each of the types offered. We examine each of these subjects in this essay.

Perhaps the major findings of this essay concern the adequacy of plan offerings. We use spanning tests to see if the plan offerings span the space offered by eight indexes. Only 53% of 417 plans span the space defined by eight indexes. This means that, for 47% of the plans, the plan participants would be better off with additional investment choices. In fact, if these plans spanned the eight indexes, participants' average return would improve by 2.3% per year, which is 21.8% of the return on an eight-index portfolio with the same level of risk. While significant on a 1-year basis, over a 20-year-period (a reasonable investment horizon for a plan participant), the cost of not offering sufficient choices makes a difference in terminal wealth of over 53%. Since, for more than one-half of plan participants, a 401(k) plan represents the participant's sole financial asset, the consequences are serious.

In the second part of this essay, we examine how well plan administrators select funds, given that they have decided to offer funds of a certain ICDI investment objective classification.

We find that on an average plan administrators select funds that underperform passive portfolios with the same risk but outperform randomly selected funds from the same ICDI category. Plan administrators show less skill in replacing or adding funds. Managers add funds that have performed well in the past and drop funds that have performed poorly. In addition, the funds they add are in ICDI categories that have performed well in the past relative to other ICDI categories (hot sectors). However, after the plans make a change, the preponderance of evidence is that added funds did no better than dropped funds. While these statements hold, on an average there do seem to be differences in skills at selecting funds by plan administrators; past performance of plans predicts future performance. The principal predictive power is with the poorer performing plans.

We also examine, whether in the aggregate plan participants allocate investments in a manner that leads to better performance than simply investing equally in each fund offered. We find no evidence to support participant allocations being superior to equal investment allocations. Participants in the aggregate show no skill in differentiating among investment choices. Thus the principal factor affecting the performance of participant 401(k) portfolios is the set of investment choices offered by plan administrators.

References

Agnew J, Balduzzi P, Sunden A (2003) Portfolio choice and trading in a large 401(k) plan. Am Econ Rev 93 193–205

Benartzi S., Thaler R (2001) Naïve diversification strategies in retirement saving plans. Am Econ Rev 91(1) 78–98

Benartzi S, Thaler R, Utkus S, Sunstein C (2007) Market rationality and legal reform. J Law Econ 50 45–79.

Blake CR, Elton EJ, Gruber MJ (1993) The performance of bond mutual funds. J Bus

Chen Z, Knez Peter (1996) Portfolio performance measures: theory and application. Review of Financial Studies 9 511–556

DeRoon F, Nijman TE, Werker BJM (2001) Testing for mean variance spanning with short sales constraints and transaction costs: the case of emerging markets. J Finance 56(2) 721–742

Elton EJ, Gruber MJ, Blake CR (2001) A first look at the accuracy of the CRSP mutual fund database and a comparison of the CRSP and Morningstar mutual fund databases. J Finance 56(6)

Elton EJ, Gruber MJ, Blake CR (2006) The adequacy of investment choices offered by 401(k) plans. J Publ Econ 90 1299–1314

Elton EJ, Gruber MJ, Blake CR (2007) Participant reaction and the performance of mutual funds offered by 401(k) plans. Journal of Financial Intermediation June

Elton EJ, Gruber MJ Busse JA (2004a) Are investors rational? Choices among index funds. J Finance 59(1) 261–288

Elton EJ, Gruber MJ, Clifton Green T (2004b), The impact of mutual fund family membership on investor risk. Unpublished manuscript, New York University

Fama E, French K (1995) Size and book-to-market factors in earnings and returns. J Finance 50 131–156

Glen J, Jorion P (1993) Currency hedging for international portfolios. J Finance 48 1965–1886

Grinblatt M, Titman S (1989) Mutual fund performance: an analysis of quarterly portfolio holdings. J Bus 39 393–411

Hansen L, Heaton J, Luttmer E (1995) Econometric evaluation of asset pricing models. Rev Finance Stud 8 237–274

Huberman G, Kandel S (1987) Mean-variance spanning. J Finance 42 873–888

Huberman G, Sengmuller P (2004) Performance and employer stock in 401(k) plans. Rev Finance vol 8

Jensen M (1968) The performance of mutual funds in the period 1945–1964. J Finance 23 389–416

Liang N, Weisbenner S (2002) Investor behavior and the purchase of company stock in 401(k) plans – the importance of plan design. Unpublished manuscript, University of Illinois

Markowitz HM (1952) Portfolio Selection. J Finance 7 no. 1, 77–91

Markowitz HM Portfolio Selection: Efficient Diversification of Investments. New York: Wiley, (1959)

Markowitz HM (1956) The optimization of a quadratic function subject to linear constraints. Naval Res. Logistics Quart 3 nos. 1, 2, March–June 111–133

Madrian B, Shea D (2001) The power of suggestion: inertia in 401(k) participants' savings. QJ Econ 116(4) 1149–1187

Mitchell O, Utkus S, Yang T (2005b) Better plans for the better paid: determinants and effects of 401(k) plan design. Pension Research Council working paper 2005

Sharpe WF (1966) Mutual fund performance. J Bus 39 119–138

Wermers R (2000) The value of active mutual fund management. J Financ Quant Anal

Chapter 15
Global Equity Risk Modeling

Jose Menchero, Andrei Morozov, and Peter Shepard

15.1 Introduction

The pioneering work of Markowitz (1952) formally established the intrinsic tradeoff between risk and return. This paradigm provided the foundation upon which the modern theory of finance was built and has proven so resilient that it has survived essentially intact for over 50 years. Almost as remarkable is the vigor with which the theory has been embraced by academics and practitioners alike.

The specific problem addressed by Markowitz was how to construct an efficient portfolio from a collection of risky assets. Markowitz defined an efficient portfolio as one that had the highest expected return for given level of risk, which he measured as standard deviation of portfolio returns. Markowitz showed that the relevant risk of an asset is not its stand-alone volatility, but rather its contribution to *portfolio* risk. Henceforth, the concepts of risk and correlation became inseparable.

A plot of expected return versus volatility for the set of all efficient portfolios maps out a curve known as the efficient frontier. In order to construct the efficient frontier using the Markowitz prescription, an investor must provide expected returns and covariances for the universe of all investable assets. The Markowitz procedure identifies the optimal portfolio corresponding to the risk tolerance of any given investor.

Tobin (1958) took the Markowitz methodology and extended it in a very simple way that nonetheless had profound implications for portfolio management. By including cash in the universe of investable assets, Tobin showed that there existed a *single* portfolio on the efficient frontier that, when combined with cash, dominated all other portfolios. For any investor, therefore, the optimal portfolio would always consist of a combination of cash and the "super-efficient" portfolio. For instance, risk-averse investors may combine the super-efficient portfolio with a large cash position, whereas risk seekers would borrow cash to purchase more of the super-efficient portfolio. Therefore, according to Tobin, the optimal investment strategy

J. Menchero (✉), A. Morozov, and P. Shepard
MSCI Barra 2100 Milvia Street, Berkley, CA 94704, USA
e-mail: jose.menchero@mscibarra.com

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_15,
© Springer Science+Business Media, LLC 2010

consists of two separate steps. The first is to determine the super-efficient portfolio. The second step is to determine the appropriate level of cash that matches the overall risk tolerance of the investor. This two-step investment process came to be known as the Tobin separation theorem.

The next major step in the development of Capital Market Theory was due to Sharpe (1964). By making certain assumptions, such as that all investors followed mean-variance preferences and agreed on the expected returns and covariances of all assets, he was able to show that the super-efficient portfolio was the market portfolio itself. Sharpe's theory, known as the Capital Asset Pricing Model, predicts that the expected return of an asset depends only on the expected return of the market and the beta of the asset relative to the market. In other words, within CAPM, the only "priced" factor is the market factor.

Using the CAPM framework, the return of any asset can be decomposed into a systematic component that is perfectly correlated with the market, and a residual component that is uncorrelated with the market. The CAPM predicts that the expected value of the residual return is zero. This does not preclude the possibility, however, of correlations among the residual returns. That is, even under the CAPM, there may be multiple sources of equity return co-movement, even if there is only one source of expected return.

Rosenberg (1974) was the first to develop multi-factor risk models to estimate the asset covariance matrix. This work was later extended by Rosenberg and Marathe (1975), who conducted a sweeping econometric analysis of multi-factor models. The intuition behind these models is that there exists a relatively parsimonious set of pervasive factors that drive asset returns. For instance, equity returns may be explained by country and industry membership, as well as by the exposure to style factors such as Size or Momentum. Returns that cannot be explained by the factors are deemed "stock specific" and are assumed to be uncorrelated. Rosenberg founded a firm, Barra, which made widespread the use of multi-factor risk models and dedicated itself to helping practitioners implement the theoretical insights of Markowitz, Tobin, Sharpe, and others. The first multi-factor risk model for the US market, dubbed USE1, was released in 1975.

The availability of rigorous multi-factor risk models such as USE1 enabled practitioners for the first time to apply the Markowitz procedure under realistic covariance assumptions. Even then, in the early days of portfolio optimization, practitioners were hampered by limitations in computational power. Consequently, much of the early research focused on the development of clever algorithms designed to overcome these obstacles, as described, for example, by Rudd and Rosenberg (1979).

Another major milestone on the path to practical implementation of modern portfolio theory came in 1989 with the development of the first Barra global equity risk model, dubbed GEM. This model was estimated via monthly cross-sectional regressions using countries, industries, and styles as explanatory factors, as described by Grinold et al. (1989).

In this paper, we present the latest Barra global equity risk model, GEM2. This model incorporates several advances and innovations over previous Barra global equity risk models. For instance, GEM2 employs an updated industry factor structure based on the *Global Industry Classification Standard* (GICS®), thus better reflecting the global economy's current industrial structure. The list of style factors has been expanded to capture additional sources of equity return co-movement. The global factor structure has been placed on an intuitive foundation with the introduction of a World factor, thereby treating industries and countries on an equal footing. GEM2 also utilizes higher frequency observations to allow for more accurate and timely risk forecasts. The GEM2 specific risk model is based on a structural approach that incorporates the latest advances in Barra risk model methodology. Finally, GEM2 uses a very broad estimation universe based on the MSCI *Global Investable Market Indices* (GIMI) for greater precision in factor return estimates.

GEM2 is available in short-term (GEM2S) and long-term (GEM2L) versions. Both versions share the same factor structure and have identical factor returns. They differ in their responsiveness. GEM2S is designed to provide users with the most accurate and responsive risk forecasts at the 1-month prediction horizon. GEM2L, by contrast, is intended for longer-term investors who value more stable risk forecasts. This paper describes only the short-term version, GEM2S. For notational efficiency, however, in this paper, we refer to this model as simply the GEM2 model.

The remainder of this paper is organized as follows. In Section 15.2, we provide an overview of the basic calculations involved in global equity risk forecasting. Section 15.3 describes the GEM2 factor structure. Section 15.4 describes estimation of the GEM2 factor returns. Section 15.5 discusses construction of the factor covariance matrix. Section 15.6 presents the specific risk model, followed by an evaluation of the accuracy of the risk forecasts in Section 15.7. This paper also contains two technical appendices. Appendix A provides details of the individual descriptors that comprise the style factors. Appendix B gives a review of bias statistics.

15.2 Forecasting Global Equity Portfolio Risk

Global investors derive return from two basic sources: price appreciation in the local currency of the asset, and repatriation of the asset value back to the base currency (numeraire) of the investor.

For concreteness, consider an investment from the perspective of a US portfolio manager. Let $C(n)$ denote the local currency of stock n, and let $r_{FX}^{C(n),\$}$ be the return of currency $C(n)$ due to exchange-rate fluctuations against the dollar. The return of stock n in US dollars, denoted $\hat{R}_n^{\$}$, is given by the usual expression

$$\hat{R}_n^{\$} = \hat{r}_n + r_{FX}^{C(n),\$} + \hat{r}_n r_{FX}^{C(n),\$}, \tag{15.1}$$

where $\hat{r}_n$ is the local return of the asset. The cross term, $\hat{r}_n \; r_{FX}^{C(n),\$}$, is typically minute and can be safely ignored for risk purposes.[1]

[1] A notable exception to this is during times of extreme inflation, when the cross term can become significant.

Our primary interest is to forecast the volatility of *excess* returns (i.e., above the risk-free rate). If $r_f^{\$}$ denotes the risk-free rate of US dollars, then $R_n^{\$} = \hat{R}_n^{\$} - r_f^{\$}$ is the excess return of stock n measured in US dollars.

Suppressing the cross term, and utilizing (15.1), we decompose the excess return into an equity component r_n and a currency component $q_n^{\$}$:

$$R_n^{\$} = r_n + q_n^{\$}. \tag{15.2}$$

The equity component is the *local excess return* of the stock,

$$r_n = \hat{r}_n - r_f^{C(n)}, \tag{15.3}$$

where $r_f^{C(n)}$ denotes the local risk-free rate for stock n. Note that r_n is independent of numeraire, so that it is the same for all global investors.

The currency component in (15.2) is given by

$$q_n^{\$} = r_f^{C(n)} + r_{FX}^{C(n),\$} - r_f^{\$}. \tag{15.4}$$

This term represents the excess return in US dollars because of holding cash denominated in the local currency of the stock.

We adopt a multi-factor framework to explain the local excess returns. This approach yields valuable insight into the underlying sources of portfolio return by separating systematic effects from the purely stock-specific component that can be diversified away. More specifically, we posit that the local excess returns are driven by a relatively small number K_E, of global equity factors, plus an idiosyncratic component unique to the particular stock,

$$r_n = \sum_{k=1}^{K_E} X_{nk} f_k + u_n. \tag{15.5}$$

Here, X_{nk} $(k \leq K_E)$ is the exposure of stock n to equity factor k, f_k is the factor return, and u_n is the specific return of the stock. The specific returns u_n are assumed to be uncorrelated with the factor returns. The factor exposures are known at the start of each period, and the factor returns are estimated via cross-sectional regression.

Suppose that there are K_C currencies in the model. Ordering the currencies after the equity factors, we can express the excess currency returns as

$$q_n^{\$} = \sum_{k=K_E+1}^{K_E+K_C} X_{nk} f_k^{\$}, \tag{15.6}$$

where X_{nk} $(k > K_E)$ is the exposure of stock n to currency k, and $f_k^{\$}$ is the excess return of the currency with respect to US dollars (which we calculate from risk-free rates and exchange rates, as in (15.4)). We take the currency exposures X_{nk} to be equal to 1 if k corresponds to the local currency of stock n, and 0 otherwise.

Equations (15.2), (15.5), and (15.6) can be combined to obtain

$$R_n^{\$} = \sum_{k=1}^{K} X_{nk} f_k^{\$} + u_n, \tag{15.7}$$

where $K = K_E + K_C$ is the total number of combined equity and currency factors. The elements of X_{nk} define the $N \times K$ factor exposure matrix, where N is the total number of stocks. Note that the factor exposure matrix is independent of numeraire.

Our treatment thus far has considered only a single asset. Our primary objective, however, is to forecast portfolio risk. The portfolio excess return (in US dollars) is given by

$$R_P^{\$} = \sum_{n=1}^{N} h_n^P R_n^{\$}, \tag{15.8}$$

where h_n^P is the portfolio weight of asset n. Note that, in general, the assets include both stocks and cash. The portfolio exposure to factor k is given by the weighted average of the asset exposures,

$$X_k^P = \sum_{n=1}^{N} h_n^P X_{nk}. \tag{15.9}$$

In order to estimate portfolio risk, we also require the factor covariance matrix and the specific risk forecasts. The elements of the factor covariance matrix are

$$F_{kl}^{\$} = \operatorname{cov}\left(f_k^{\$}, f_l^{\$}\right), \tag{15.10}$$

where the dollar superscript indicates that at least some of the matrix elements depend on the base currency of the investor. Note, however, that the $K_E \times K_E$ block of the factor covariance matrix corresponding to equity factors is independent of numeraire.

The covariance of specific returns is

$$\Delta_{mn} = \operatorname{cov}(u_m, u_n). \tag{15.11}$$

For most stocks, we assume that the specific returns are uncorrelated, so that the off-diagonal elements of Δ_{mn} are zero. However, (15.11) represents the generalized case in which the specific returns of some securities are linked.[2]

[2] We relax the assumption of uncorrelated specific returns for different share classes of the same stock.

The portfolio risk in US dollars can now be obtained as the square root of the variance,

$$\sigma(R^{\$}_{P}) = \left[\sum_{kl} X_k^P F_{kl}^{\$} X_l^P + \sum_{mn} h_m^P \Delta_{mn} h_n^P \right]^{1/2}. \tag{15.12}$$

Although (15.12) is written for portfolio risk, it is equally valid for tracking error by replacing portfolio weights and exposures with their active counterparts.

Many global investors, of course, are interested in risk forecasts from different numeraire perspectives. The only term on the right-hand-side of (15.12) that depends on base currency is the factor covariance element, $F_{kl}^{\$}$. However, These elements can be transformed to some other numeraire, γ. Substituting these transformed factor covariances F_{kl}^{γ} into (15.12) yields risk forecasts with respect to the new base currency γ.

15.3 Factor Exposures

15.3.1 Estimation Universe

The coverage universe is the set of all securities for which the model provides risk forecasts. The estimation universe, by contrast, is the subset of stocks that is used to estimate the model. Judicious selection of the estimation universe is a critical component in building a sound risk model. The estimation universe must be sufficiently broad to accurately represent the investment opportunity set of global investors, without being so broad as to include illiquid stocks that may introduce spurious return relationships into the model. Furthermore, the estimation universe should be reasonably stable to ensure that factor exposures are well behaved across time. *Representation, liquidity*, and *stability*, therefore, represent the three primary goals that must be attained when selecting a risk model estimation universe.

A well-constructed equity index must address and overcome these very issues and therefore serves as an excellent foundation for the estimation universe. The GEM2 estimation universe utilizes the MSCI *All Country World Investable Market Index* (ACWI IMI), part of the MSCI Global Investable Market Indices family which represents the latest in MSCI index-construction methodology. MSCI ACWI IMI aims to reflect the full breadth of global investment opportunities by targeting 99% of the float-adjusted market capitalization in 48 developed and emerging markets. The index-construction methodology applies innovative rules designed to achieve index stability, while reflecting the evolving equity markets in a timely fashion. Moreover, liquidity screening rules are applied to ensure that only investable stocks with reliable pricing are included for index membership.

If a country is deemed excessively restrictive to foreign investment, it is excluded from MSCI ACWI IMI. Nevertheless, many of these frontier markets represent important opportunities for certain global investors. For instance, since 2002, Qualified Foreign Institutional Investors (QFIIs) are permitted to invest directly in the

Chinese domestic A-share market, which does not constitute part of MSCI ACWI IMI. Similarly, the six Gulf Cooperation Council (GCC) markets (Saudi Arabia, Oman, Qatar, UAE, Kuwait, and Bahrain) are important to many global investors, though they are also excluded from MSCI ACWI IMI. We therefore supplement the estimation universe with stocks from China Domestic and the six GCC countries. We down-weight these markets in the regression in order to not distort the factor relationships within the investable universe.

In June 2002, the MSCI Global Investable Market Indices instituted the transition to float-adjusted market-cap weighting and the application of additional liquidity screening rules. In this month, the number of constituents in MSCI ACWI IMI dropped sharply from nearly 12,000 to below 7,000. Although this corresponds to a drop of only 5% on a cap-weighted basis, this event must be treated carefully in the estimation universe to avoid potentially spurious jumps in factor exposures. We remove the discontinuity in the estimation universe by backward-excluding all stocks that were dropped from MSCI ACWI IMI in June 2002. We allow exceptions to this rule in cases where additional stocks are needed to populate countries with few assets.

The official MSCI ACWI IMI history begins in June 1994. To obtain a deeper model history, however, we backward-extend the estimation universe for 2 years by including the constituents of the standard MSCI *All Country World Index*, which contains large-cap and mid-cap stocks. We supplement this universe with all US and Japanese small-cap stocks with MSCI ACWI IMI membership as of June 1994. While this does lead to a jump in the number of estimation universe constituents in June 1994, it does not lead to jumps in risk forecasts since it occurs before the first available factor covariance matrix.

15.3.2 GEM2 Factor Structure

The equity factor set in GEM2 includes a World factor (w), countries (c), industries (i), and styles (s). Every stock is assigned an exposure of 1 to the World factor. Hence, the local excess returns in (15.5) can be rewritten as

$$r_n = f_w + \sum_c X_{nc} f_c + \sum_i X_{ni} f_i + \sum_s X_{ns} f_s + u_n. \tag{15.13}$$

Mathematically, the World factor represents the intercept term in the cross-sectional regression. Economically, it describes the aggregate up-and-down movement of the global equity market. Typically, the World factor is the dominant source of risk for a diversified long-only portfolio.

For most institutional investors, however, the primary concern is the risk of *active* long/short portfolios. If both the portfolio and benchmark are fully invested – as is typically the case – then the active exposure to the World factor is zero. Similarly, if the long and short positions of a long/short portfolio are of equal absolute value, the net exposure to the World factor is zero. Thus, we must look beyond the World factor to other sources of risk.

Country factors play a critical role in global equity risk modeling. One reason is that they are powerful indicator variables for explaining the cross section of global equity returns. A second, related, reason is that the country allocation decision is central to many global investment strategies, and portfolio managers often must carefully monitor their exposures to these factors. We therefore include explicit country factors for all markets covered.

In Table 15.1, we present a list of the 55 countries covered by GEM2, together with their corresponding currencies. The country exposures X_{nc} in GEM2 are set equal to 1 if stock n is in country c, and set equal to 0 otherwise. We assign country exposures based on country membership within the MSCI ACWI IMI, MSCI China A Index and MSCI GCC Countries Index. Note that depository receipts and cross-listed assets are assigned factor exposures for the underlying or primary asset, as defined by the MSCI Equity Indices.

We also show in Table 15.1 the average and ending weights (based on total market capitalization) over the period, from January 1997 to January 2008. It is interesting to note the relative decline of the US and Japanese markets, and the rise of markets such as Brazil, Russia, India, and China. We also report the number of securities in the estimation universe for each country and the market capitalization (float-adjusted and total) in billions of US Dollars as of January 2008.

Industries are also important variables in explaining the sources of global equity return co-movement. One of the major strengths of GEM2 is to employ the Global Industry Classification Standard (GICS®) for the industry factor structure. The GICS scheme is hierarchical, with 10 top-level sectors, which are then divided into 24 industry groups, 68 industries, and 154 sub-industries. GICS applies a consistent global methodology to classify stocks based on careful evaluation of the firm's business model and economic operating environment. The GICS structure is reviewed annually by MSCI Barra and Standard & Poor's to ensure that it remains timely and accurate.

Identifying which industry factors to include in the model involves a combination of judgment and empirical analysis. At one extreme, we could use the 10 GICS sectors as industry factors. Such broad groupings, however, would certainly fail to capture much of the cross-sectional variation in stock returns. At the other extreme, we could use all 154 sub-industries as the factor structure. Besides the obvious difficulties associated with the unwieldy numbers of factors (e.g., risk reporting, thin industries), such an approach would present a more serious problem for risk forecasting: although adding more factors always increases the in-sample R^2 of the cross-sectional regressions, many of the factor returns would not be statistically significant. Allowing noise-dominated "factors" into the model defeats the very purpose of a factor risk model.

In GEM2, selection of the industry factor structure begins at the second level of the GICS hierarchy, with each of the 24 industry groups automatically qualifying as a factor. This provides a reasonable level of granularity, without introducing an excessive number of factors. We then analyze each industry group by carefully examining the industries and sub-industries contained therein to determine if a more granular factor structure is warranted. The basic criteria, we use to guide

Table 15.1 GEM2 country factors and currencies

Country code	Country name	Currency name	Average weight	Jan-08 weight	Jan-08 # stocks	Jan-08 float-cap	Jan-08 total-cap
ARG	Argentina	Argentine Peso	0.10	0.09	13	19.43	41.67
AUS	Australia	Australian Dollar	1.62	2.48	236	1045.32	1197.26
AUT	Austria	Euro	0.17	0.37	35	94.68	176.79
BHR	Bahrain (*)	Bahraini Dinar	0.01	0.02	11	3.13	10.40
BEL	Belgium	Euro	0.64	0.69	47	195.07	333.81
BRA	Brazil	Brazilian Real	0.65	1.87	140	531.80	904.92
CAN	Canada	Canadian Dollar	2.59	3.44	297	1372.74	1658.38
CHL	Chile	Chilean Peso	0.18	0.26	34	48.63	126.19
CHN	China Domestic (*)	Chinese Yuan	1.90	7.00	1511	1209.86	3379.20
CHX	China International	Hong Kong Dollar	0.67	2.96	215	637.21	1429.87
COL	Colombia	Colombian Peso	0.03	0.07	11	13.38	32.18
CZE	Czech Republic	Czech Koruna	0.06	0.15	8	28.85	74.13
DNK	Denmark	Danish Krone	0.39	0.46	57	153.77	220.54
EGY	Egypt	Egyptian Pound	0.05	0.17	31	33.44	81.72
FIN	Finland	Euro	0.59	0.69	52	281.29	334.69
FRA	France	Euro	4.18	5.24	184	1611.22	2528.75
DEU	Germany	Euro	3.31	3.71	184	1431.50	1790.54
GRC	Greece	Euro	0.28	0.48	62	136.76	229.30
HKG	Hong Kong	Hong Kong Dollar	1.07	1.50	165	395.80	723.88
HUN	Hungary	Hungarian Forint	0.06	0.09	8	29.21	41.89
IND	India	Indian Rupee	0.54	2.41	251	382.40	1164.89
IDN	Indonesia	Indonesian Rupiah	0.13	0.34	54	74.93	164.14
IRE	Ireland	Euro	0.26	0.27	30	112.05	128.76
ISR	Israel	Israeli Shekel	0.20	0.32	69	89.30	152.88
ITA	Italy	Euro	2.05	2.09	161	633.60	1009.46
JPN	Japan	Japanese Yen	11.52	8.96	1160	3198.33	4322.65
JOR	Jordan	Jordanian Dinar	0.03	0.05	17	4.43	23.73
KOR	Korea	Korean Won	1.02	2.06	332	604.46	993.49

(continued)

Table 15.1 (continued)

Country code	Country name	Currency name	Average weight	Jan-08 weight	Jan-08 # stocks	Jan-08 float-cap	Jan-08 total-cap
KWT	Kuwait (*)	Kuwaiti Dinar	0.14	0.33	85	72.41	157.53
MYS	Malaysia	Malaysian Ringgit	0.43	0.53	119	108.21	255.87
MEX	Mexico	Mexican Peso	0.43	0.59	40	172.93	285.52
MAR	Morocco	Moroccan Dirham	0.03	0.09	10	10.40	45.13
NLD	Netherlands	Euro	1.47	1.34	70	445.70	645.96
NZL	New Zealand	New Zealand Dollar	0.09	0.07	22	24.77	33.57
NOR	Norway	Norwegian Krone	0.34	0.71	68	186.40	344.12
OMN	Oman (*)	Omani Rial	0.01	0.03	14	5.92	15.84
PAK	Pakistan	Pakistan Rupee	0.03	0.06	24	8.35	29.97
PER	Peru	Peruvian Sol	0.04	0.11	16	25.06	54.67
PHL	Philippines	Philippine Peso	0.06	0.14	27	21.78	67.20
POL	Poland	Polish Zloty	0.11	0.33	66	73.29	157.75
PRT	Portugal	Euro	0.20	0.24	21	57.70	117.57
QAT	Qatar (*)	Qatari Rial	0.05	0.17	23	17.25	83.92
RUS	Russia	Russian Ruble	0.49	2.83	71	306.42	1364.70
SAU	Saudi Arabia (*)	Saudi Rial	0.32	0.75	40	169.49	363.13
SGP	Singapore	Singapore Dollar	0.48	0.72	104	186.03	348.62
ZAF	South Africa	South African Rand	0.63	0.86	116	293.06	415.28
ESP	Spain	Euro	1.41	2.11	88	664.97	1018.36
SWE	Sweden	Swedish Krone	1.05	1.05	105	389.92	506.61
CHE	Switzerland	Swiss Franc	2.54	2.41	118	1021.22	1163.85
TWN	Taiwan	Taiwan Dollar	1.25	1.34	411	436.94	646.55
THA	Thailand	Thailand Bhat	0.19	0.36	66	59.72	174.41
TUR	Turkey	New Turkish Lira	0.18	0.45	68	72.55	214.91
GBR	UK	U.K. Pound	8.65	7.65	463	3456.81	3694.66
ARE	UAE (*)	Emirati Dirham	0.06	0.43	40	44.77	206.47
USA	US	US Dollar	47.53	34.80	2468	15742.12	16794.26
Total			102.50	108.74	10138	38446.78	52482.54

Weights are normalized excluding China Domestic and GCC (marked by asterisk), and computed within the GEM2 estimation universe using total market capitalization. Average weights are from January 1997 to January 2008. Market capitalizations are reported in billions of US Dollars industry factor selection, are: (a) the groupings of industries into factors must be economically intuitive, (b) the industry factors should have a strong degree of statistical significance, (c) incorporating an additional industry factor should significantly increase the explanatory power of the model, and (d) thin industries (those with few assets) should be avoided.

The result of this process is the set of 34 GEM2 industry factors, presented in Table 15.2. Industries that qualify as factors tend to exhibit volatile returns and have significant weight. We find that this relatively parsimonious set of factors captures

Table 15.2 GEM2 Industry Factors

GICS sector	GEM2 code	GEM2 industry factor name	Average weight	Jan-08 weight
Energy	1	Energy equipment & services	0.75	1.29
	2	Oil, gas & consumable fuels	4.88	9.32
	3	Oil & gas exploration & production	1.00	1.72
Materials	4	Chemicals	2.36	2.84
	5	Construction, containers, paper	1.38	1.24
	6	Aluminum, diversified metals	1.05	2.41
	7	Gold, precious metals	0.37	0.58
	8	Steel	0.79	1.83
Industrials	9	Capital goods	7.33	8.60
	10	Commercial & professional services	1.43	0.77
	11	Transportation Non-Airline	1.82	2.32
	12	Airlines	0.37	0.45
Consumer discretionary	13	Automobiles & components	2.52	2.29
	14	Consumer durables & apparel	2.33	1.93
	15	Consumer services	1.35	1.39
	16	Media	3.24	2.11
	17	Retailing	3.42	2.08
Consumer staples	18	Food & staples retailing	1.82	1.76
	19	Food, beverage & tobacco	4.56	4.37
	20	Household & personal products	1.43	1.20
Health care	21	Health care equipment & services	2.13	1.93
	22	Biotechnology	0.78	0.68
	23	Pharmaceuticals, life sciences	6.17	3.82
Financials	24	Banks	10.52	10.83
	25	Diversified financials	5.63	5.06
	26	Insurance	4.61	4.14
	27	Real estate	2.08	3.07
Information Technology	28	Internet software & services	0.62	0.74
	29	IT services, software	3.24	2.56
	30	Communications equipment	2.46	1.41
	31	Computers, electronics	3.69	2.81
	32	Semiconductors	2.47	1.52
Telecom	33	Telecommunication services	7.11	5.84
Utilities	34	Utilities	4.31	5.08

Weights are computed within the GEM2 estimation universe using total market capitalization. Average weights are computed from January 1997 to January 2008 most of the in-sample R^2 explained by the 154 sub-industries, but with a much higher degree of statistical significance. Also reported in Table 15.2 are the average and end-of-period industry weights from January 1997 to January 2008. The weights were computed using the entire GEM2 estimation universe (i.e., including China Domestic and GCC countries). Only five industries have end-of-period weights less than 100 bps, and these tend to be highly volatile, thus making them useful risk factors.

Investment style represents another important source of systematic risk. Style factors, also known as *risk indices*, are designed to capture these sources of risk. They are constructed from financially intuitive stock attributes called *descriptors*, which serve as effective predictors of equity return covariance. Since the descriptors within a particular style factor are meant to capture the same underlying driver of returns, these descriptors tend to be significantly collinear. For instance, price-to-book ratio, dividend yield, and earnings yield are all attributes used to identify value stocks, and they tend to exhibit significant cross-sectional correlation. Although these descriptors have significant explanatory power on their own, naively including them as separate factors in the model may lead to serious multi-collinearity problems. Combining these descriptors into a single style factor overcomes this difficulty and also leads to a more parsimonious factor structure.

Unlike country and industry factors, which are assigned exposures of either 0 or 1, style factor exposures are continuously distributed. To facilitate comparison across style factors, they are standardized to have a mean of 0 and a standard deviation of 1. Each descriptor is also standardized similarly. That is, if d_{nl}^{Raw} is the raw value of stock n for descriptor l, then the standardized descriptor value is given by

$$d_{nl} = \frac{d_{nl}^{Raw} - \mu_l}{\sigma_l}, \tag{15.14}$$

where μ_l is the cap-weighted mean of the descriptor (within the estimation universe), and σ_l is the equal-weighted standard deviation. We adopt the convention of standardizing using the cap-weighted mean so that a well-diversified cap-weighted global portfolio, such as MSCI ACWI IMI, has approximately zero exposure to all style factors. For the standard deviation, however, we use the equal-weighted mean to prevent large-cap stocks from having an undue influence on the overall scale of the exposures.

Formally, descriptors are combined into risk indices as follows

$$X_{nk} = \sum_{l \in k} w_l \, d_{nl}, \tag{15.15}$$

where w_l is the descriptor weight, and the sum takes place over all descriptors within a particular risk index. Descriptor weights are determined using an optimization algorithm to maximize the explanatory power of the model. Style factor exposures are rescaled to have a standard deviation of 1.

Some of the style factors are standardized on a *global-relative* basis, others on a *country-relative* basis. In the former case, the mean and standard deviation in (15.14) are computed using the entire global cross section. In the latter case, the factors have mean 0 and standard deviation 1 within each country. When deciding which standardization convention to adopt, we consider both the intuitive meaning of the factor and its explanatory power.

GEM2 uses eight style factors. Below, we provide a qualitative description of each of the style factors:

- The *Volatility* factor is typically the most significant style factor. In essence, it captures market risk that cannot be explained by the World factor. The most important descriptor within the Volatility index is historical beta relative to the World portfolio (as proxied by the estimation universe). To better understand this factor, consider a fully invested long-only portfolio that is strongly tilted toward high-beta stocks. Intuitively, this portfolio has greater market risk than a portfolio with beta equal to one. This additional market risk is captured through positive exposure to the Volatility factor. Note that the time-series correlation between the World factor and the Volatility factor is typically very high, so that these two sources of risk add coherently in this example. If, by contrast, the portfolio is invested in low-beta stocks, then the risk from the Volatility and the World factors is partially canceled, as intuitively expected. We standardize the Volatility factor on a global-relative basis. As a result, the mean exposure to Volatility within a country can deviate significantly from zero. This standardization convention is a natural one for a global model, as most investors regard stocks in highly volatile markets as having more exposure to the factor than those in low-volatility markets. This view is reflected in the data, as we find that the explanatory power of the factor is greater using the global-relative standardization.
- The *Momentum* factor often ranks second in significance after Volatility. Momentum differentiates stocks based on recent relative performance. Descriptors within Momentum include historical alpha from a 104-week regression and relative strength (over trailing 6 and 12 months) with a 1-month lag. Similarly to Volatility, Momentum is standardized on a global-relative basis. This is also an intuitive convention for a global model. From the perspective of a global investor, a stock that strongly outperforms the World portfolio is likely to be considered a positive momentum stock, even if it slightly underperforms its country peers. The empirical results support this view, as the Momentum factor standardized globally has greater explanatory power than one standardized on a country-relative basis.
- The *Size* factor represents another well-known source of return covariance. It captures the effect of large-cap stocks moving differently from small-cap stocks. We measure Size by a single descriptor: log of market capitalization. The explanatory power of the model is quite similar whether Size is standardized globally or on a country-by-country basis. We adopt the country-relative standardization, however, as it is more intuitive and consistent with investors' perception of the markets. For instance, major global equity indices, such as the MSCI Global Investable Market Indices, segment each country according to size, with the largest stocks inside each country always being classified as large-cap stocks. Moreover, standardizing the Size factor on a global-relative basis would serve as an unintended proxy for developed markets versus emerging markets, and increases collinearity with the country factors.

- The *Value* factor describes a major investment style which seeks to identify stocks that are priced low relative to fundamentals. We standardize Value on a country-relative basis. This again is consistent with the way major indices segment each market, with each country divided roughly equally into value and growth sub-indices. This convention also circumvents the difficulty of comparing fundamental data across countries with different accounting standards. GEM2 utilizes official MSCI data items for Value factor descriptors, as described in the *MSCI Barra Fundamental Data Methodology Handbook* (2008).
- *Growth* differentiates stocks based on their prospects for sales or earnings growth. It is standardized on a country-relative basis, consistent with the construction of the MSCI Value and Growth Indices. Therefore, each country has approximately half the weight in stocks with positive Growth exposure, and half with negative exposure. The GEM2 Growth descriptors also utilize official MSCI data items, as described in the *MSCI Barra Fundamental Data Methodology* handbook.
- The *Non-Linear Size (NLS)* factor captures non-linearities in the payoff to size exposure across the market-cap spectrum. NLS is based on a single raw descriptor: the cube of the log of market capitalization. Since this raw descriptor is highly collinear with the Size factor, we orthogonalize it to the Size factor. This procedure does not affect the fit of the model, but does mitigate the confounding effects of collinearity and thus preserves an intuitive meaning for the Size factor. The NLS factor is represented by a portfolio that goes long mid-cap stocks and shorts large-cap and small-cap stocks.
- The *Liquidity* factor describes return patterns to stocks based on relative trading activity. Stocks with high turnover have positive exposure to Liquidity, whereas low-turnover stocks have negative exposure. Liquidity is standardized on a country-relative basis.
- *Leverage* captures the return difference between high-leverage and low-leverage stocks. The descriptors within Leverage include market leverage, book leverage, and debt-to-assets ratio. This factor is standardized on a country-relative basis.

In Appendix A, we provide additional details on the individual descriptors comprising each style factor.

15.4 Factor Returns

15.4.1 Data Quality and Outlier Treatment

The most difficult and time-consuming part of constructing a global equity risk model lies in the preparation of the input data. If the data inputs are garbage, the risk forecasts will likewise be garbage, no matter how sophisticated the model. Assuring a high degree of data quality therefore is a vital part of building a reliable risk model.

In order to obtain the highest quality inputs, GEM2 leverages the same data infrastructure used for the construction of MSCI Global Investable Market Indices.

Data items such as raw descriptors, GICS codes, country classifications, and clean daily stock returns are obtained from in-house sources that have already undergone extensive quality control.

No matter how stringent the data quality assurance process, we can never exclude the possibility of extreme outliers entering the data set. These extreme outliers may represent legitimate values or outright data errors. Either way, such observations must be carefully handled to prevent a few data points from having an undue impact on model estimation.

In GEM2, we employ a multi-step algorithm to identify and treat outliers. The algorithm assigns each observation into one of three groups. The first group represents values so extreme that they are treated as potential data errors and removed from the estimation process. The second group represents values that are regarded as legitimate, but nonetheless so large that their impact on the model must be limited. We *trim* these observations to three standard deviations from the mean. The third group of observations, which forms the bulk of the distribution, consists of values less than three standard deviations from the mean. These observations are left unadjusted.

The thorough data quality assurance process, coupled with the robust outlier algorithm, ensures that the input data are clean and reliable. There is still the issue, however, of dealing with *missing* data. Data could be missing either due to the lack of availability or due to the removal by the outlier algorithm. The data items most likely to be missed include descriptors for the Value, Growth, Leverage, and Liquidity style factors. More rarely, stocks may be missing exposure to Volatility and Momentum factors; this occurs for securities which lack return history, such as recent spin-offs or IPOs. Other data items, such as Size, industry, and country exposures, are never missing, since these items are required for coverage universe membership.

If the underlying data required to compute factor exposures are missing, then the factor exposures are generated using a data-replacement algorithm. A simple approach would be to assign zero exposure to all missing values. A better approach, however, is to exploit known relationships between factor exposures. For instance, if we know that a stock is in the Semiconductor industry and has a positive exposure to Volatility, then it is likely that the stock also has a positive exposure to Growth.

We apply the data-replacement algorithm to the Value, Growth, Leverage, and Liquidity factors. The algorithm works by regressing these four factors (using only non-missing exposures) against Size, Volatility and industries. The slope coefficients of the regression are then used to estimate the factor exposures for the stocks with missing data. Countries are not used as explanatory variables in the regression since the four style factors (Value, Growth, Leverage and Liquidity) are standardized to be mean zero with respect to countries. The algorithm is not applied to the other four style factors (Size, NonLinear Size, Volatility and Momentum) since these factor exposures are rarely, if ever, missing. Note also that the algorithm is applied at the factor level as opposed to the descriptor level. In other words, if a stock has data for some descriptors within a risk index, but not others, then the non-missing data will be used to compute the factor exposures. Only if *all* descriptors are absent does the replacement algorithm become active.

15.4.2 GEM2 Regression Methodology

The equity factor returns f_k in GEM2 are estimated by regressing the local excess returns r_n against the factor exposures X_{nk},

$$r_n = \sum_{k=1}^{K_E} X_{nk} f_k + u_n. \tag{15.16}$$

GEM2 uses weighted least squares, assuming that the variance of specific returns is inversely proportional to the square root of total market capitalization.

As described in Section 15.3, the GEM2 equity factors include the World factor, countries, industries, and styles. Every stock in GEM2 has unit exposure to the World factor, and indicator variable exposures of 0 or 1 to countries and industries. As a result, the sum of all country factors equals the World factor, and similarly for industries, i.e.,

$$\sum_c X_{nc} = 1, \text{ and } \sum_i X_{ni} = 1, \tag{15.17}$$

for all stocks n. In other words, the sum of all country columns in the factor exposure matrix gives a column with 1 in every entry, which corresponds to the World factor. The same holds for industry factors. The GEM2 factor structure therefore exhibits exact twofold collinearity. Constraints must be applied to obtain a unique solution.

In GEM2, we adopt constraints as in Heston and Rouwenhorst (1994) that require the cap-weighted country and industry factor returns to sum to zero,

$$\sum_c w_c f_c = 0, \quad \text{and} \quad \sum_i w_i f_i = 0, \tag{15.18}$$

where w_c is the weight of the estimation universe in country c, and w_i is the corresponding weight in industry i. These constraints remove the exact collinearities from the factor exposure matrix, without reducing the explanatory power of the model.

We can now give a more precise interpretation to the factors. Consider the cap-weighted estimation universe, with holdings h_n^E. The return of this portfolio R_E can be attributed using the GEM2 factors,

$$R_E = f_w + \sum_c w_c f_c + \sum_i w_i f_i + \sum_s X_s^E f_s + \sum_n h_n^E u_n. \tag{15.19}$$

The constraints imply that the first two sums in (15.19) are equal to zero. The third sum is also zero since the style factors are standardized to be cap-weighted mean zero; i.e., $X_s^E = 0$, for all styles s. The final sum in (15.19) corresponds to the specific return of a broadly diversified portfolio and is *approximately* zero (note it would be *exactly* zero if we used regression weights instead of capitalization weights). Thus, to an excellent approximation, (15.19) reduces to

$$R_E \approx f_w. \tag{15.20}$$

In other words, the return of the World factor is essentially the cap-weighted return of the estimation universe.

To better understand the meaning of the pure factor portfolios, we report in Table 15.3 the long and short weights (as of January 2008) of the World portfolio and several pure factor portfolios in various market segments. The World portfolio is represented by the cap-weighted GEM2 estimation universe. The pure World factor is 100% net long and the net weights closely match those of the World portfolio in each segment. The other pure factors have net weight of zero and therefore represent long/short portfolios.

As a first approximation, the pure country factors can be regarded as going long 100% the particular country, and going short 100% the World portfolio. For instance, going long 100% Japan and short 100% the World portfolio results in a portfolio with roughly 91% weight in Japan, and −91% in all other countries. The pure country factors, however, have zero exposure to industry factors. This is accomplished by taking appropriate long/short combinations in other countries. For instance, the Japanese market is over-represented in the segment corresponding to the Automobile factor. To partially hedge this exposure, the pure Japan factor takes a net short position of −1.09% in the US Automobile segment. A similar short position would be found in the German Automobile segment.

The pure Automobile factor can be thought, as a first approximation, to be formed by going 100% long the Automobile industry and 100% short the World portfolio. A more refined view of the factor takes into account that the net weight in each country is zero. The pure Automobile factor naturally takes a large long position in Japanese automobiles, but hedges the Japan exposure by taking short positions in other Japanese segments.

The pure Volatility factor is perhaps the easiest to understand, as it takes offsetting long and short positions within all segments corresponding to GEM2 factors (e.g., Japan, US, and Automobiles). Note that the weights are not equal to zero for segments that do not correspond to GEM2 factors, such as Japanese automobiles.

15.4.3 Characteristics of GEM2 Factors

In this section, we present and discuss some of the quantitative characteristics of the GEM2 factors. In particular, we investigate the degree of collinearity among the factor exposures and report on the statistical significance, performance, and volatility of the factor returns.

A feature of the multi-factor framework is that it disentangles the effects of many variables acting simultaneously. Multi-collinearity among factors, however, can confound this clean separation of effects.

One measure of collinearity is the pair-wise cross-sectional correlation between factor exposures. In Table 15.4, we report regression-weighted correlations among style factors and industries, averaged over the period from January

Table 15.3 Segment weights, as of January 2008, for the World portfolio and several pure factor portfolios

Segment	(ESTU) World portfolio	Pure World factor	Pure Japan factor	Pure US factor	Pure Auto factor	Pure Volatility factor
World (Net)	100.00	100.00	0.00	0.00	0.00	0.00
Long	100.00	106.61	97.09	66.41	112.61	54.91
Short	0.00	−6.61	−97.09	−66.41	−112.61	−54.91
Japan (Net)	8.96	8.96	91.04	−8.96	0.00	0.00
Long	8.96	10.00	91.04	0.18	33.22	5.79
Short	0.00	−1.05	0.00	−9.14	−33.22	−5.79
US (Net)	34.80	34.80	−34.80	65.20	0.00	0.00
Long	34.80	34.99	1.55	65.21	19.59	16.37
Short	0.00	−0.19	−36.34	0.00	−19.59	−16.37
Auto (Net)	2.33	2.33	0.00	0.00	97.67	0.00
Long	2.33	2.45	4.31	0.96	97.67	1.16
Short	0.00	−0.12	−4.31	−0.96	0.00	−1.16
Japan Auto (Net)	1.05	0.87	4.31	−0.06	33.21	0.18
Long	1.05	0.89	4.31	0.05	33.21	0.38
Short	0.00	−0.03	0.00	−0.11	0.00	−0.20
US Auto (Net)	0.21	0.27	−1.09	0.89	15.51	0.22
Long	0.21	0.27	0.00	0.89	15.51	0.32
Short	0.00	0.00	−1.09	0.00	0.00	−0.10

The World portfolio is represented by the cap-weighted GEM2 estimation universe

1997 to June 2008. In general, the correlations are intuitive in sign. For instance, Volatility is positively correlated with Growth, Liquidity, and Semiconductors, and negatively with Value and Utilities. Value, as expected, is positively correlated with Banks and Utilities, and negatively with Growth and Biotechnology. Although none of the correlations are particularly large, the correlations between pairs of style factors are typically larger, on average, than those between styles and industries.

The statistical significance of factor returns plays a key role in the construction of a risk model. Let f_k be the factor return for a particular period, and $\mathrm{se}(k)$ be the standard error of this estimate. The t-statistic is given by

$$t_k = \frac{f_k}{\mathrm{se}(k)}. \tag{15.21}$$

Generally, a t-statistic with absolute value greater than 2 is considered significant at the 95% confidence level. Ideally, a risk factor should have high t-statistics that are persistent across time. Useful measures of this are the average squared t-statistic over time, and the percentage of observations with significant t-statistics.

In Table 15.5, we report the summary statistics for the GEM2 style factors. In this case, we divide the sample period into two sub-periods. In the first sample period (from January 1997 to June 2002), the style factors were generally more volatile and had a greater degree of statistical significance than in the second sub-period (July 2002 to June 2008). This is not too surprising, since the first sub-period contains the rise and fall of the internet bubble. Note that the kurtosis of the Value factor in the second sample period is particularly high. This is caused primarily by 2 six-sigma events in August 2007.

We can discern from Table 15.5 a clear risk hierarchy in the style factors. Alone at the top is the Volatility factor, which is far more volatile and significant than any of the other factors. In the second tier are Momentum and Size, which are clearly set apart from the third group, containing the other five styles. In this third group, the Value factor tends to be the most significant.

Performance, however, tells a different story. Volatility performed very poorly during the first sub-period and was essentially flat during the second. Value and Momentum, on the other hand, were strongly positive over both sample periods.

There is substantial correlation between some of the style factors with the estimation universe. For instance, from Table 15.5, we see that Volatility is strongly correlated (about 80%) with the estimation universe. This makes sense since during up-markets, high-beta stocks tend to outperform low-beta stocks, so that the return to Volatility – which includes beta – is also positive. A similar argument holds for down-markets. Liquidity and Size also tend to be positively correlated with the estimation universe over both sample periods.

Although not reported here, the country and industry factor returns are highly significant. The mean average squared t-statistic is 31.7 for countries and 17.2 for industries over the sample period from January 1997 to June 2008. Also, more than half of the observations are statistically significant at the 95% confidence level.

Table 15.4 Regression-weighted cross-sectional correlation of style and industry factor exposures

Factor	Volatility	Momentum	Size	Value	Growth	NL Size	Liquidity	Leverage
Volatility	1.00	−0.05	−0.09	−0.25	0.27	−0.06	0.42	−0.07
Momentum	−0.05	1.00	0.12	−0.17	0.08	0.15	0.09	−0.08
Size	−0.09	0.12	1.00	−0.11	−0.05	0.11	0.10	−0.03
Value	−0.25	−0.17	−0.11	1.00	−0.24	−0.03	−0.15	0.22
Growth	0.27	0.08	−0.05	−0.24	1.00	−0.02	0.18	−0.17
Non-linear Size	−0.06	0.15	0.11	−0.03	−0.02	1.00	0.11	0.01
Liquidity	0.42	0.09	0.10	−0.15	0.18	0.11	1.00	−0.03
Leverage	−0.07	−0.08	−0.03	0.22	−0.17	0.01	−0.03	1.00
Energy equipment & services	0.06	0.02	−0.03	−0.02	0.06	0.02	0.07	−0.03
Oil, gas & consumable fuels	−0.02	0.02	0.11	0.06	−0.03	−0.04	0.00	−0.01
Oil & gas exploration & production	0.01	0.02	−0.01	0.01	0.00	0.00	0.02	0.01
Chemicals	−0.03	0.00	−0.01	0.04	−0.04	0.03	−0.01	0.00
Construction, containers, paper	−0.04	−0.01	−0.02	0.07	−0.06	0.04	−0.04	0.06
Aluminum, diversified metals	0.03	0.02	0.02	0.03	0.00	0.01	0.03	0.00
Gold, precious metals	0.00	−0.01	0.00	−0.06	0.00	0.00	−0.01	−0.04
Steel	0.04	0.02	−0.01	0.09	−0.02	0.01	0.03	0.03
Capital goods	0.01	0.01	−0.07	0.04	−0.02	0.01	0.00	−0.01
Commercial & professional services	0.01	0.00	−0.09	−0.05	0.06	−0.01	−0.02	−0.03
Transportation non-airline	−0.06	0.01	0.00	0.02	−0.02	0.02	−0.04	0.08
Airlines	0.02	−0.02	0.01	0.02	−0.01	0.02	0.02	0.08
Automobiles & components	0.01	0.00	0.01	0.10	−0.02	0.00	−0.01	0.00
Consumer durables & apparel	0.01	−0.01	−0.07	0.05	0.00	0.00	0.03	−0.06
Consumer services	−0.02	0.00	−0.06	−0.03	0.04	0.02	0.01	0.02
Media	−0.01	−0.03	0.01	−0.12	0.02	0.02	−0.03	0.00
Retailing	0.03	0.01	−0.05	−0.01	0.05	0.01	0.04	−0.06
Food & staples retailing	−0.07	−0.01	0.02	−0.02	0.00	0.01	−0.03	−0.01

(continued)

Table 15.4 (continued)

Factor	Volatility	Momentum	Size	Value	Growth	NL Size	Liquidity	Leverage
Food, beverage & tobacco	−0.14	0.00	0.02	−0.01	−0.05	0.02	−0.07	−0.01
Household & personal products	−0.05	0.00	0.03	−0.04	−0.02	−0.01	−0.03	−0.03
Health care equipment & services	−0.02	0.02	−0.06	−0.08	0.09	0.01	0.04	−0.05
Biotechnology	0.09	0.00	−0.05	−0.17	0.07	−0.03	0.07	−0.05
Pharmaceuticals, life sciences	−0.04	0.00	0.06	−0.10	0.02	−0.04	0.00	−0.10
Banks	−0.08	0.00	0.12	0.12	−0.09	−0.05	−0.10	0.10
Diversified financials	0.05	0.00	0.02	0.04	0.01	0.01	−0.03	0.04
Insurance	−0.06	0.00	0.04	0.11	−0.05	0.03	−0.06	−0.10
Real estate	−0.13	0.02	−0.09	0.07	−0.08	0.03	−0.07	0.17
Internet software & services	0.15	−0.02	−0.04	−0.12	0.12	−0.03	0.07	−0.06
IT services, software	0.15	−0.01	−0.06	−0.14	0.12	−0.02	0.09	−0.12
Communications equipment	0.16	−0.02	−0.01	−0.09	0.06	−0.03	0.08	−0.07
Computers, electronics	0.11	−0.02	−0.02	−0.05	0.04	−0.03	0.09	−0.09
Semiconductors	0.21	−0.02	0.00	−0.10	0.09	−0.02	0.16	−0.08
Telecommunication services	0.04	−0.02	0.15	−0.07	0.01	−0.07	0.01	0.05
Utilities	−0.17	0.01	0.06	0.11	−0.13	0.04	−0.07	0.20

Results are averages over the period January 1997 to June 2008. Correlations above 0.10 in absolute value are shaded in gray

Table 15.5 Style factor summary statistics based on weekly factor returns

Factor name	Average squared t-statistic	Percent observations with $\|t\| > 2$	Factor kurtosis (weekly)	Annualized factor return	Annualized factor volatility	Factor Sharpe ratio	Correlation with ESTU
A. *January 1997 to June 2002 (66 months)*							
Volatility	119.46	82.6	4.92	−5.98	8.21	−0.73	0.79
Momentum	48.12	78.5	6.34	7.62	4.46	1.71	0.11
Size	43.65	79.9	4.73	2.37	2.97	0.80	0.29
Value	11.76	57.6	4.55	7.65	1.96	3.89	−0.16
Growth	9.75	50.7	4.49	−0.09	1.61	−0.05	0.41
Non-linear Size	8.86	49.7	4.31	1.61	1.89	0.85	0.15
Liquidity	10.18	53.8	4.00	4.99	1.63	3.06	0.51
Leverage	4.73	33.0	5.55	−1.18	1.13	−1.04	−0.03
Average	32.06	60.72	4.86	2.12	2.98	1.06	0.26
B. *July 2002 to June 2008 (72 months)*							
Volatility	89.50	84.8	4.07	0.36	4.98	0.07	0.84
Momentum	35.37	76.4	5.71	4.62	2.65	1.75	−0.11
Size	26.45	65.0	5.33	0.71	1.67	0.42	0.41
Value	11.89	53.1	10.74	3.86	1.37	2.83	0.08
Growth	4.45	30.4	4.86	1.00	0.80	1.24	0.04
Non-linear Size	6.85	44.0	5.34	1.25	1.11	1.13	0.17
Liquidity	8.46	46.3	4.34	0.35	1.25	0.28	0.46
Leverage	4.30	33.7	8.59	−0.09	0.84	−0.11	0.02
Average	23.41	54.21	6.12	1.51	1.83	0.95	0.24

The sample period, January 1997 to June 2008, is divided into two sub-periods, A and B

Finally, it is worth noting that the time-series correlation between the World factor and the cap-weighted estimation universe return is 99.4%, thus confirming our interpretation of the World factor.

15.4.4 Comparing the Explanatory Power of the Models

An important summary statistic used to measure the explanatory power of a model is the R^2 of the cross-sectional regressions,

$$R^2 = 1 - \frac{\sum_n w_n\, u_n^2}{\sum_n w_n\, r_n^2}, \tag{15.22}$$

where w_n is the regression weight of stock n, u_n is the specific return, and r_n is the local excess return. The R^2 can be interpreted as the ratio of the average *explained* squared return to the average *total* squared return.

In Fig. 15.1, we plot the trailing 12-month R^2 of GEM2 factor groups over time. The sample period is from January 1994 to June 2008. The first case we examine is the World factor in isolation. This is the single most important factor and, on average, explains about 10% in R^2. Its explanatory power, however, fluctuates over time. For instance, in 2000 the World factor explained only a few percentage points, whereas this figure rose to nearly 20% by 2003.

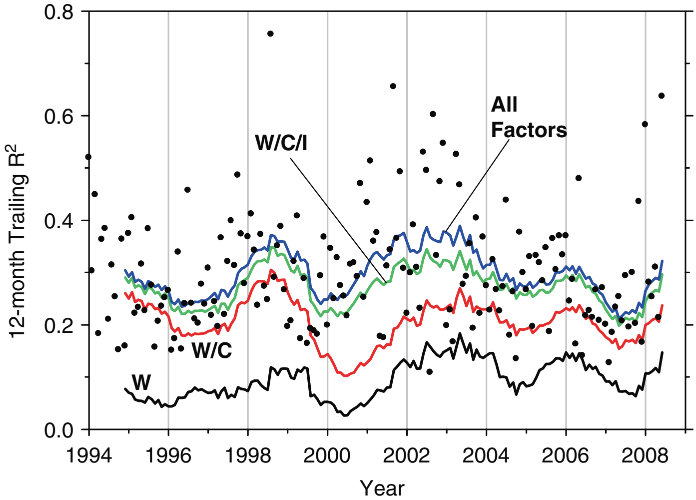


Fig. 15.1 Explanatory power of GEM2 factor groups, expressed as trailing 12-month R^2. Here, *W* indicates the World factor in isolation, *W/C* denotes World factor plus countries, and *W/C/I* represents World, country, and industry factors. The *black dots* indicate the actual monthly R^2 values. Over the sample period January 1994 to June 2008, the average R^2 values are approximately 10% (W), 21% (W/C), 27% (W/C/I) and 30% (all factors)

From Fig. 15.1, it is evident that most of the explanatory power of the model comes from the combined effect of country and industry factors. Interestingly, the *relative* explanatory power of these factor groups changes significantly over time. In the early period, industries explained relatively little compared with countries. Then, some time in 1999, the explanatory power of industries suddenly exploded. From 2000 to 2003, industries actually explained a greater proportion of trailing 12-month R^2 than countries. Since 2003, the situation has partially reversed, with country factors reasserting their dominance though industry factors remained very important. The average R^2 obtained using the World factor with country factors is about 21%, whereas adding industry factors on top of the countries increases the R^2 to about 27%. In other words, country factors add about 11% in R^2 above the World factors, and industries add another 6%.[3] This example illustrates some the insights obtained by the GEM2 factor structure, which places countries and industries on an equal footing.

The style factors contribute about 300 bps in R^2 beyond that explained by countries and industries alone. This amount, however, also varies considerably over time. In the early period, style factors added only about 100 bps, whereas this widened to about 500 bps during the 2000–2004 period. More recently, style factors have added about 200 bps.

The black dots in Fig. 15.1 represent the monthly R^2 values for the complete set of GEM2 factors. This quantity varies dramatically from month to month, and tends to be largest during big market moves. In extreme months, commonalities in returns swamp asset-specific returns. For example, the largest single R^2 observed during the sample period was 0.75. This occurred in August 1998, when the market dropped by more than 15%.

15.5 GEM2 Factor Covariance Matrix

The GEM2 factor covariance matrix is built from the time series of weekly factor returns, f_{kt}. Weekly observations reduce sampling error and provide more responsive risk forecasts than monthly observations. Daily factor returns are not used for computing global covariances due to asynchronous trading effects, but they are computed for performance attribution purposes.

We use exponentially weighted moving averages (EWMA) to estimate the factor covariance matrix for both equity and currency blocks. This approach gives more weight to recent observations and is a simple, yet robust, method for dealing with data non-stationary. An alternative approach would be to use generalized auto-regressive conditional heteroskedasticity, or GARCH. We find that EWMA estimators are typically more robust to changing market dynamics and produce more accurate risk forecasts than their GARCH counterparts.

[3] We obtain similar results if industry factors are introduced before country factors.

In GEM2, we must also account for the possibility of serial correlation in factor returns, which can affect risk forecasts over a longer horizon. Suppose, for instance, that high-frequency returns are negatively correlated. In this case, long-horizon risk forecasts estimated on high-frequency data will be lower than that implied using simple square-root-of-time scaling, since returns one period tend to be partially offset by opposing returns the following period.

The prediction horizon in Barra risk models is typically 1 month. Models that are estimated on daily or weekly returns therefore must adjust for serial correlation. Note that models estimated on monthly observations need not adjust for serial correlation, since the observation frequency coincides with the prediction horizon.

Full treatment of serial correlation must account for not only the correlation of one factor with itself across time, but also for the correlation of two factors with each other across different time periods. In GEM2, we model serial correlation using the methodology of Newey and West (1987) with two lags. This assumes that the return of any factor may be correlated with the return of any other factor up to 2 weeks prior.

It is useful to think of the factor covariance matrix as being composed of a correlation matrix, which is then scaled by the factor volatilities. Volatilities and correlations can then be estimated separately using EWMA with different half-life parameters.

The volatility half-life determines the overall responsiveness of the model. Selecting the proper volatility half-life involves analyzing the trade-off between accuracy and responsiveness on the one hand, and stability on the other. If the volatility half-life is too long, the model gives undue weight to distant observations that have little to do with current market conditions. This leads to stable risk forecasts but at the cost of reduced accuracy. By contrast, a short volatility half-life makes the model more responsive, generally improving the bias statistics, but at the cost of jumpier risk forecasts. Of course, the volatility half-life cannot be lowered without bound; if the half-life is too short, then sampling error can become so large that the risk forecasts are not only less stable, but also less accurate.

We measure the accuracy of the risk forecasts using bias statistics, which essentially represent the ratio of realized risk to forecast risk. A bias statistic of 1 is considered ideal, but sampling error ensures that realized bias statistics will deviate from 1, even for perfect risk forecasts. In Appendix B, we provide a review of bias statistics.

When carrying out bias statistic testing, one must consider the number of months to include in the observation window. Sampling error suggests that one should use the largest interval possible, since that would minimize sampling error. This would be a compelling argument for stationary data. However, In reality financial markets are non-stationary. It is possible to over-predict risk of several years and then under-predict for others, while obtaining a bias statistic close to 1. Getting the right *average* risk forecast over 10 or 20 years is small consolation to a portfolio manager who can be wiped out due to poor risk forecasts in any single year.

Rolling 12-month intervals provide a more relevant framework for evaluating the accuracy of risk forecasts. Although sampling error is larger, this approach penalizes the model for poor risk forecasts during any individual year. For a given portfolio, we compute the rolling absolute deviation (RAD) of the bias statistic as follows. First, we calculate the bias statistic over the first year of the observation window, using 12 monthly observations. We then take the absolute value of the deviation of the bias statistic from 1. We then roll the 12-month window forward 1 month at a time, computing absolute deviations each month, until we reach the end of the observation window. For instance, a 138-month observation window will accommodate 127 separate 12-month windows. The RAD for the portfolio is the average of the absolute deviations over these 127 sub-periods. The RAD measure captures in a single number the accuracy of risk forecasts over rolling 12-month intervals by penalizing all observations throughout the sample history that deviate from the ideal bias statistic of 1.

As shown in Appendix B, for perfect risk forecasts and normally distributed returns (kurtosis 3), the critical value of the RAD over a 138-month observation window is 0.22. In other words, if the observed RAD is larger than 0.22, then with 95% confidence, we can reject the null hypothesis of an accurate risk forecast. Returns, of course, are not normally distributed. For a kurtosis of 5, which corresponds to slightly fat tails, the critical value of RAD for a 138-month observation window is 0.29. For even fatter tails, say a kurtosis of 10, the critical value increases to 0.34.

To study the question of the stability of risk forecasts, we compute the monthly absolute change in volatility for each factor k,

$$v_k(t) = \frac{|\sigma_k(t+1) - \sigma_k(t)|}{\sigma_k(t)}. \tag{15.23}$$

If the volatility forecast changes over 1 month from 6.0 to 5.7%, for instance, then the monthly absolute change is 5%.

The *mean variability* provides a measure of the stability of the risk forecasts and is given by the average of the monthly absolute changes,

$$\bar{v} = \frac{1}{KT} \sum_{k,t} v_k(t), \tag{15.24}$$

where K is the number of factors and T is the number of time periods.

In Fig. 15.2, we show the mean RAD for all GEM2 equity factors, plotted as a function of volatility half-life. The bias statistics were computed over rolling 12-month windows using a 138-month observation window from January 1997 to June 2008. We also plot the mean monthly variability $\bar{v}$ of the equity factors. The volatility half-life of the GEM2 model is indicated by the vertical line.

The solid lines in Fig. 15.2 are the results which incorporate a two-lag autocorrelation adjustment, as in GEM2. For comparison, we also show the corresponding results by dashed lines without serial correlation adjustments. Including autocorrelation adjustments makes the risk forecasts more accurate but also slightly increases the monthly variability in the forecasts.

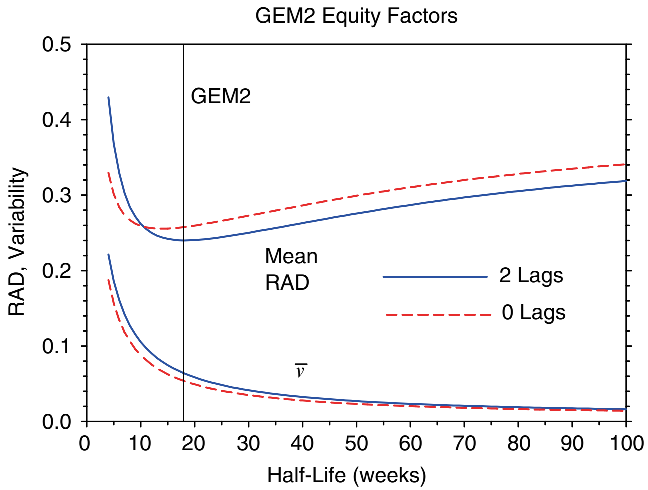


Fig. 15.2 Mean 12-month rolling absolute deviation (RAD) and mean Variability ($\bar{v}$) of GEM2 equity factors, plotted as a function of volatility half-life. The sample period was 138 months from January 1997 to June 2008. The solid lines indicate results with the two-lag autocorrelation that is used GEM2, and the dashed lines are for no autocorrelation adjustment. The vertical line indicates the actual GEM2 half-life parameter

From Fig. 15.2, we see that the volatility half-life of GEM2 (18 weeks) minimizes the mean RAD and thus produces the most accurate risk forecasts. Some users, however, place a premium on the *stability* of the risk forecast. In this case, a longer half-life may be desirable even though the risk forecasts are slightly less accurate. For instance a 52-week volatility half-life produces a mean RAD of about 0.27 (vs. 0.24 for GEM2), but the monthly variability is reduced to 2.5% (vs. 6.7% for the 18-week half-life). This example illustrates the tradeoff between accuracy and stability.

Figure 15.2 also clearly demonstrates the effect of sampling error. Namely, as the half-life is reduced below about 15 weeks, the RAD increases dramatically while the variability of the risk forecast also increases.

We also investigated how the accuracy and stability of *currency* risk forecasts varied with volatility half-life. Although not shown here, the results are qualitatively and quantitatively almost identical to the results reported in Fig. 15.2. Namely, the 18-week volatility half-life minimizes the mean RAD (at 0.26), and produces a mean variability of 6.6%. A half-life of 52 weeks produces an RAD of 0.30 but has more stable risk forecasts with a mean variability of 2.6%. The effect of autocorrelation for currencies is also comparable to that shown in Fig. 15.2.

Special care must be exercised when selecting the correlation half-life. To ensure a well-conditioned correlation matrix, the number of observations T must be significantly greater than the number of factors K. In the extreme case that $T < K$, the correlation matrix will have zero eigenvalues. This can present problems for portfolio optimization, as it would be possible to find spurious active portfolios with seemingly zero factor risk.

For GEM2, we use a correlation half-life of 104 weeks. A useful rule of thumb is that the number of effective observations is roughly three times the half-life. Therefore, a correlation half-life of 104 weeks corresponds to roughly 300 effective observations. The volatility and correlation half-lives used in GEM2 are the same as those used in other Barra single-country models, such as USE3S.

Another complication in estimating the factor covariance matrix arises for the case of missing factor returns. One possible reason for missing data is holidays. For instance, stocks do not trade in China during Golden Week, so it is not possible to directly estimate a return to the China Domestic factor. More commonly, however, missing factor returns arise from using time series of differing lengths. For instance, the GCC stocks appear in the estimation universe as of July 2000, so the factor returns prior to this date are missing.

We use the EM algorithm of Dempster, Laird, and Rubin (1977) to estimate the factor covariance matrix in the presence of missing factor returns. This method employs an iterative procedure to estimate the covariance matrix, conditional on all available data. The EM algorithm assures that the factor covariance matrix is positively semi-definite.

15.6 Specific Risk

15.6.1 GEM2 Specific Risk Model

GEM2 uses a structural model to estimate specific risk. This methodology is similar to that used in other Barra risk models, but employs weekly rather than monthly observations. These higher-frequency observations allow us to make the forecasts more responsive while also reducing estimation error.

Mathematically, the specific risk forecast of a stock is given as the product of three components

$$\sigma_n = \hat{S}\left(1 + \hat{V}_n\right) K_M, \tag{15.25}$$

where $\hat{S}$ is the forecast mean absolute specific return of stocks in the estimation universe, $\hat{V}_n$ is the forecast relative absolute specific return, and K_M is a multiplicative scalar which depends on the market-cap segment M of the stock. The overall responsiveness of the model is dominated by $\hat{S}$, whereas $\hat{V}_n$ determines the cross-sectional variation in specific risk. The role of K_M is to convert from weekly absolute specific returns to monthly standard deviation. The motivation for modeling specific risk in terms of absolute specific returns is that the estimation process is less prone to outliers.

The *realized* mean absolute specific return for a given period t is computed as the cap-weighted average,

$$S_t = \sum_n w_{nt}\,|u_{nt}|, \tag{15.26}$$

where u_{nt} is the realized specific return for stock n. We use the cap-weighted average because large-cap assets are less prone to extreme outliers. The forecast mean absolute specific return is given by the exponentially weighted average of the realized values,

$$\hat{S} = \sum_t \gamma_t S_t. \tag{15.27}$$

For GEM2, we use a half-life parameter of 9 weeks for γ_t. To determine this half-life parameter, we first constructed a broad group of active portfolios. We then selected the half-life parameter γ_t so that the specific risk forecast variability of these portfolios roughly matched the variability in forecasts of the factors, while also taking into account model performance. This results in highly accurate specific risk forecasts and greater stability in the proportion of risk coming from factors and stock-specific components.

In Fig. 15.3, we plot the realized mean weekly absolute specific return over the period from June 1992 to June 2008, together with the forecast values from GEM2. Overall, the model does an excellent job forecasting the absolute specific return levels. In the mid-1990s, the absolute specific returns were quite low (about 2%), but then began to rise, hitting a peak of nearly 6% in early 2000. After the collapse of the internet bubble, absolute specific return levels fell sharply over the next 5 years. Since 2007, however, specific risk levels have again begun to rise.

We use a factor model to forecast the relative absolute specific return, whose *realized* values are given by

$$\varepsilon_{nt} = \frac{|u_{nt}| - S_t}{S_t}. \tag{15.28}$$

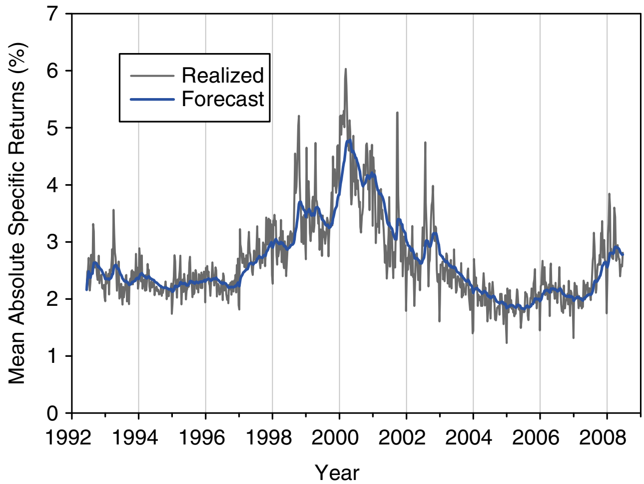


Fig. 15.3 Plot of weekly realized and forecast mean absolute specific returns for the GEM2 estimation universe

GEM2 factors already provide a sound basis for explaining ε_{nt}, since the level of relative absolute specific risk depends on the country, industry, and style exposures of the individual stocks. For instance, large-cap utility stocks tend to have lower specific returns than, say, small-cap internet stocks.

While the model factors are well suited to forecast specific risk, the most powerful explanatory variable is simply the trailing realized specific volatility. We therefore augment the GEM2 factor exposure matrix with this additional factor for the purpose of forecasting the relative absolute specific returns.

To estimate the relationship between factor exposures and relative absolute specific returns, we perform a 104-week pooled cross-sectional regression,

$$\varepsilon_{nt} = \sum_k \tilde{X}_{nk}^t g_k + \lambda_{nt}, \tag{15.29}$$

where $\tilde{X}_{nk}^t$ represents the augmented factor exposure matrix element. We use capitalization weights in the regression to reduce the impact of high-kurtosis small-cap stocks and apply exponential weighting with a half-life of 26 weeks. The forecast relative absolute specific return is given by

$$\hat{V}_n = \sum_k \tilde{X}_{nk} g_k, \tag{15.30}$$

where $\tilde{X}_{nk}$ is the most recent factor exposure matrix, and g_k is the slope coefficient estimated over the trailing 104 weeks.

The product $\hat{S}\ (1 + \hat{V}_n)$ is the weekly forecast absolute specific return for the stock. We need to convert this to a monthly standard deviation. If the specific returns were normally distributed (temporally), then the time-series standard deviation of the ratio

$$\frac{u_{nt}}{S_t(1 + \hat{V}_{nt})} \tag{15.31}$$

would be $\sqrt{\pi/2}$, where $\hat{V}_{nt}$ is the forecast relative absolute specific return at time t. However, since the specific returns exhibit kurtosis, the standard deviation is usually slightly greater than this theoretical value. For this reason, the multiplicative scalar K_M is sometimes called the *kurtosis correction*.

There is one final adjustment that must be applied to the kurtosis correction. The prediction horizon of the risk model is 1 month. We observe, on average, a small but persistent negative serial correlation in weekly specific returns. As a result, the monthly specific volatility will be slightly less than that suggested by simple square root of time scaling. If we rank stocks by degree of serial correlation, however, we find no relationship between the rank of stocks over non-overlapping periods. We therefore define the standardized return with two-lag autocorrelation,

$$b_{nt} = \frac{u_{n,t} + u_{n,t-1} + u_{n,t-2}}{\sqrt{3} S_t(1 + \hat{V}_{nt})}. \tag{15.32}$$

The kurtosis correction is given by the realized standard deviation of b_{nt},

$$K_M = \sigma(b_{nt}), \tag{15.33}$$

where the standard deviation is computed over all observations within a particular market-cap segment M over the trailing 104 weeks.

15.6.2 Specific Risk Bias Statistics

In this section, we evaluate the accuracy of the specific risk forecasts. We consider the entire 138-month sample period and compute rolling 12-month bias statistics for each stock. There are 127 rolling 12-month windows that fit into the sample period. The number of stocks in the estimation universe on average is approximately 8,000. Thus, we compute roughly one million 12-month bias statistics.

The cap-weighted distribution of these bias statistics is shown in Fig. 15.4. The mean of the distribution is 1.02, indicating that on average the specific risk forecasts are essentially unbiased. The cap-weighted mean RAD is approximately 0.23, which is slightly more accurate than the mean RAD of 0.24 that was reported for factor volatility forecasts (see Fig. 15.2). Approximately, 86% of the observations fall within the confidence interval. If returns were normally distributed, then we would expect about 95% of the observations to fall within the confidence interval. In fact, kurtosis values of about 5 are typical of specific returns. In Fig. 15.6 of Appendix B,

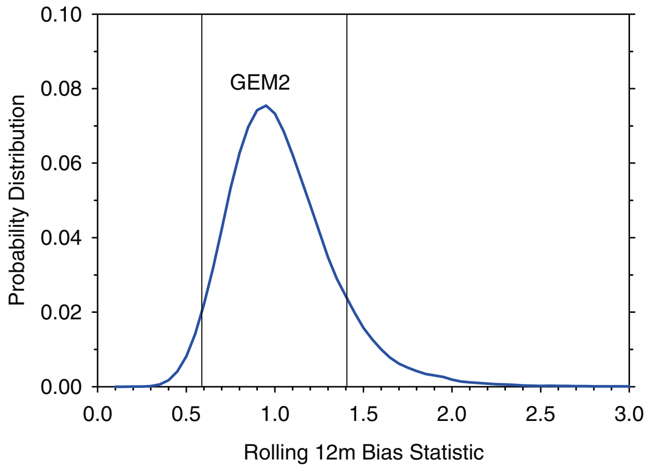


Fig. 15.4 Frequency distribution of 12-month rolling bias statistics for specific risk forecasts. Sample period is January 1997 to June 2008. Results are cap-weighted within the GEM2 estimation universe. Vertical lines indicate confidence interval [0.59, 1.41]

we show that for a kurtosis of 5 over 138 periods, about 86% of the observations should fall within the confidence interval for perfect risk forecasts, consistent with the GEM2 results.

We also computed the distribution of equal-weighted rolling 12-month bias statistics. The results are largely similar to the cap-weighted case. The equal-weighted mean bias statistic is 1.01, and about 82% of the observations fall within the confidence interval. The mean RAD is 0.26 in this case. These results indicate that the GEM2 specific risk forecasts are highly accurate across the entire market-cap spectrum.

15.7 Model Performance

In this section, we evaluate the accuracy of risk forecasts for a combination of country, industry, and style portfolios. The country and industry portfolios were constructed by carving out the constituents of the GEM2 estimation universe from the appropriate segment, and then cap-weighting the stocks. The style portfolios were constructed by ranking all stocks according to their exposures to the particular style factor, and then cap-weighting the stocks that ranked in the top and bottom 20%. For each of the long-only country, industry, and style portfolios thus formed, we also construct long/short active portfolios by selecting MSCI ACWI IMI as the benchmark. The sample period in all cases was from January 1997 to June 2008.

In Fig. 15.5, we plot the distribution of the 12-month rolling bias statistics for the complete set of country, industry, and style *active* portfolios. The sample period

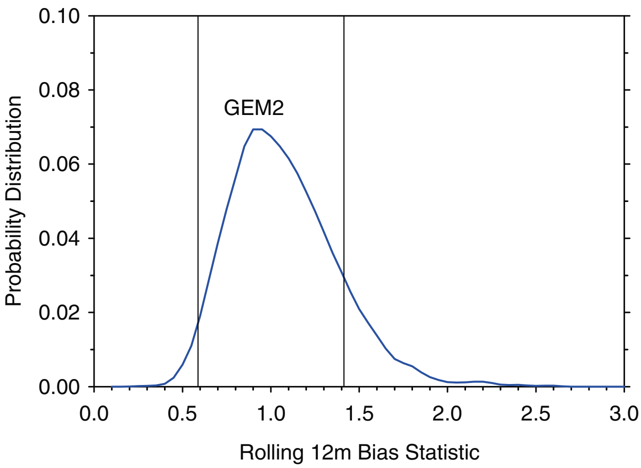


Fig. 15.5 Frequency distribution of 12-month rolling bias statistics for country, industry and style active portfolios (long/short) described in text. Sample period is January 1997 to June 2008. Vertical lines indicate the confidence interval [0.59, 1.41]

was from January 1997 to June 2008. The vertical lines denote the approximate confidence interval assuming normally distributed returns. The mean of the distribution is 1.04, with a mean RAD of 0.24. About 86% of the observations fall within the confidence interval. The kurtosis values of the active portfolios are typically in the range of 4–5. At these kurtosis levels, Fig. 15.6 of Appendix B indicates that for perfect risk forecasts about 86–90% of the observations should fall within the confidence interval, roughly consistent with the GEM2 results.

We also investigated the bias statistic distribution for the same set of active portfolios over two sub-periods. The first sub-period runs from January 1997 to June 2002 (66 months), and corresponds to a period of generally rising volatility. The second sub-period runs from July 2002 to June 2008 (72 months), and corresponds to a period of falling volatility. In the first sub-period, the mean rolling 12-month bias statistic is 1.10. For the second sub-period, it is 0.98. The fact that these results do not deviate far from 1 indicates that the GEM2 model does a good job of adapting to changing market dynamics.

15.8 Conclusion

Forecasting global equity risk is an integral part of any disciplined global investment strategy. In this paper, we presented a detailed description of the Barra global equity model, GEM2. This model incorporates the latest advances in equity risk modeling, including (a) an improved factor structure to better capture the cross section of asset returns, (b) higher frequency observations for reduced sampling error and more responsive forecasts, (c) the introduction of a World factor to cleanly separate the effects of countries and industries, and (d) a structural specific risk model for highly accurate forecasts.

Appendix A: Descriptors by Style Factor

Volatility (Global Relative)

Historical beta (β_n) (*HBETA*)

Computed as the slope coefficient in a time-series regression of local excess stock returns r_{nt} against the cap-weighted local excess return of the estimation universe R_t,

$$r_{nt} = \alpha_n + \beta_n R_t + e_{nt}. \tag{15.34}$$

The regression coefficients are estimated on the trailing 104 weeks of returns.

Daily standard deviation (*DASTD*)

This descriptor differentiates stocks based on recent volatility, and is computed as the standard deviation of daily local returns over the past 65 trading days.

Cumulative range (*CMRA*)

This descriptor differentiates stocks that have experienced wide swings over the last 12 months from those that have traded in a narrow range. Let $Z(T)$ be the cumulative local excess logarithmic return over the past T months,

$$Z(T) = \sum_{t=1}^{T} \ln(1 + r_{nt}) - \sum_{t=1}^{T} \ln(1 + r_{ft}). \tag{15.35}$$

where r_{nt} is the local return of stock n for month t, and r_{ft} is the risk-free return of the local currency. The cumulative range is given by

$$CMRA = \ln(1 + Z_{\max}) - \ln(1 + Z_{\min}), \tag{15.36}$$

where $Z_{\max} = \max\{Z(T)\}$, $Z_{\min} = \min\{Z(T)\}$, and $T = 1, \ldots, 12$.

Historical sigma (σ_n) (*HSIGMA*)

Computed as the volatility of residual returns in (15.34),

$$\sigma_n = \text{std}(e_{nt}). \tag{15.37}$$

The volatility is estimated over the trailing 104 weeks of returns.

Momentum (Global Relative)

Historical alpha (α_n) (*HALPHA*)

Given by the intercept term α_n in (15.34). Regression coefficients are estimated on trailing 104 weeks of returns.

Relative strength (6 months) (*RSTR6*)

Computed as the cumulative local excess logarithmic return over the previous 6 months,

$$RSTR6 = \sum_{t=1}^{6} \ln(1 + r_{nt}) - \sum_{t=1}^{6} \ln(1 + r_{ft}), \tag{15.38}$$

with a 1-month lag. Here, r_{nt} is the local return of stock n for month t, and r_{ft} is the risk-free return of the local currency.

Relative strength (12 months) (*RSTR12*)

Computed as the cumulative local excess logarithmic return over the previous 12 months,

$$RSTR12 = \sum_{t=1}^{12} \ln(1 + r_{nt}) - \sum_{t=1}^{12} \ln(1 + r_{ft}), \tag{15.39}$$

with a 1-month lag.

Size (Country Relative)

Log of market cap (*LNCAP*)

Given by the logarithm of the total market capitalization of the firm.

Value (Country Relative)

Predicted earnings-to-price ratio (*EPFWD*)

Given by the 12-month forward-looking earnings per share (*EPS12F*) divided by the current price. Forward-looking earnings per share are defined as a weighted average between the mean analyst-predicted earnings per share for the current and next fiscal years. For details, refer to Section 1.3 of *MSCI Barra Fundamental Data Methodology*.

Book-to-price ratio (*BTOP*)

Computed using the latest book value per share and the price. For details, refer to Section 1.2.5 of *MSCI Barra Fundamental Data Methodology*.

Cash earnings-to-price ratio (*CETOP*)

Computed using the trailing 12-month cash earnings per share (*CEPS*) divided by the current price. For details, refer to Section 1.2.5 of *MSCI Barra Fundamental Data Methodology*.

Trailing earnings-to-price ratio (*ETOP*)

Given by the trailing 12-month earnings per share (*EPS*) divided by the current price. For details on the calculation of the *EPS*, refer to Section 2.1 of *MSCI Barra Fundamental Data Methodology*.

Dividend-to-price ratio (*YIELD*)

Given by the annualized dividend-per-share (DPS) divided by the current price. For details, refer to Section 1.2.4 of *MSCI Barra Fundamental Data Methodology*.

Growth (Country Relative)

Long-term predicted earnings growth (*EGRLF*)

Long-term (3–5 years) earnings growth rate forecasted by analysts. For details, refer to Section 2.2.5 of *MSCI Barra Fundamental Data Methodology*.

Sales growth (trailing 5 years) (*SGRO*)

Annual reported sales per share are regressed against time over the past 5 fiscal years. The slope coefficient is then divided by the average annual sales per share to obtain the sales growth. For details, refer to Section 2.2.1 of *MSCI Barra Fundamental Data Methodology*.

Earnings growth (trailing 5 years) (*EGRO*)

Annual reported earnings per share are regressed against time over the past 5 fiscal years. The slope coefficient is then divided by the average annual earnings per share to obtain the earnings growth. For details, refer to Section 2.2.1 of *MSCI Barra Fundamental Data Methodology*.

Non-linear Size (Country Relative)

Cube of log of market cap (*NLSIZE*)

First, the standardized size exposure (i.e., log of market cap) is cubed. The resulting factor is then orthogonalized to the size factor on a regression-weighted basis. The orthogonalized factor is then winsorized and standardized.

Liquidity (Country Relative)

Average share turnover, trailing 12 months (*STOA*)

Computed as the logarithm

$$STOA = \log\left(\frac{1}{12}\sum_t \frac{V_t}{S_t}\right), \tag{15.40}$$

where V_t is the trading volume of the asset for month t, and S_t is the number of shares outstanding. The sum runs over the past 12 months.

Average share turnover, trailing 3 months (*STOQ*)

Computed as the logarithm

$$STOQ = \log\left(\frac{1}{3}\sum_t \frac{V_t}{S_t}\right), \tag{15.41}$$

where V_t is the trading volume of the asset for month t, and S_t is the number of shares outstanding. The sum runs over the past 3 months.

Share turnover, 1 month (*STOM*)

Computed as the logarithm

$$STOM = \log\left(\frac{V_t}{S_t}\right), \tag{15.42}$$

where V_t is the trading volume of the asset for the month, and S_t is the corresponding number of shares outstanding.

Leverage (Country Relative)

Market leverage (*MLEV*)

Computed as

$$MLEV = \frac{MCAP + PREF + LD}{MCAP}, \tag{15.43}$$

where *MCAP* is the market value of common equity at previous month-end, *PREF* is the most recent book value of preferred equity, and *LD* is the most recent book value of long-term debt. For details, refer to Section 15.5 of *MSCI Barra Fundamental Data Methodology*.

Book leverage (*BLEV*)

Computed as

$$BLEV = \frac{BV + PREF + LD}{BV}, \tag{15.44}$$

where *BV* is the most recent book value of common equity, *PREF* is the most recent book value of preferred equity, and *LD* is the most recent book value of long-term debt. For details, refer to Section 15.5 of *MSCI Barra Fundamental Data Methodology*.

Debt-to-assets (*DTOA*)

Computed as

$$DTOA = \frac{TD}{TA}, \tag{15.45}$$

where *TD* is the book value of total debt (long-term debt, *LD*, and current liabilities, *CL*), and *TA* is most recent book value of total assets. For details, refer to Section 15.5 of *MSCI Barra Fundamental Data Methodology*.

Appendix B: Review of Bias Statistics

B1. Single-Window Bias Statistics

To assess a model's predictive performance, it is not enough to consider its explanatory power; we must also test how well its risk forecasts perform out of sample. In this section, we describe how to evaluate the accuracy of risk forecasts.

A commonly used measure to assess a risk model's accuracy is the bias statistic. The bias statistic, conceptually, represents the ratio of realized risk to forecast risk.

Let r_{nt} be the return to portfolio n over period t, and let σ_{nt} be the beginning-of-period volatility forecast. Assuming perfect forecasts, the *standardized* return,

$$b_{nt} = \frac{r_{nt}}{\sigma_{nt}}, \tag{15.46}$$

has expected standard deviation 1. The bias statistic for portfolio n is the *realized* standard deviation of standardized returns,

$$B_n = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left(b_{nt} - \bar{b}_n\right)^2}, \tag{15.47}$$

where T is the number of periods in the observation window.

Assuming normally distributed returns and perfect risk forecasts, for sufficiently large T the bias statistic B_n is approximately normally distributed about 1, and roughly 95% of the observations fall within the confidence interval,

$$B_n \in \left[1 - \sqrt{2/T}, 1 + \sqrt{2/T}\right]. \tag{15.48}$$

If B_n falls outside this interval, then we reject the null hypothesis that the risk forecast was accurate.

If returns are not normally distributed, however, then fewer than 95% of the observations will fall within the confidence interval, *even for perfect risk forecasts*. In Fig. 15.6, we show simulated results for the percentage of observations actually

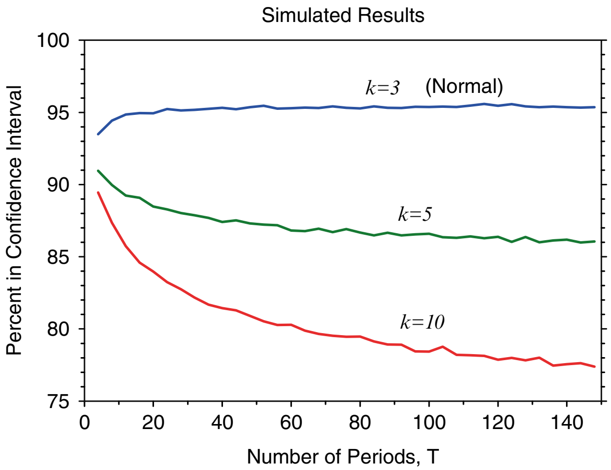


Fig. 15.6 Percent of observations falling within confidence interval $\left[1 - \sqrt{2/T}, 1 + \sqrt{2/T}\right]$, where T is the number of periods in the observation window. Results were simulated using a normal distribution ($k = 3$), and using a t-distribution with kurtosis values $k = 5$ and $k = 10$. The standard deviations were equal to 1 in all cases (i.e., perfect risk forecasts). For the normal distribution, the percentage of observations inside the confidence interval quickly approaches 95%. As kurtosis is increased, however, the proportion within the confidence interval declines considerably

falling within this interval, plotted versus observation-window length T, for several values of kurtosis k.

For the normal case (kurtosis $k = 3$), except for the smallest values of T, the confidence interval indeed captures about 95% of the observations. As the kurtosis increases, however, the percentage falling within the interval drops significantly. For instance, even for a fairly modest kurtosis level of 5, only 86% of bias statistics fall inside the confidence interval for an observation window of 120 periods.

B2. Rolling-Window Bias Statistics

The purpose of bias-statistic testing is to assess he accuracy of risk forecasts, typically over a long sample period. Let T be the length of the observation window, which corresponds to the number of months in the sample period. One possibility is to select the entire sample period as a single window, and to compute the bias statistic as in (15.47). This would be a good approach if financial data were stationary, as sampling error is reduced by increasing the length of the window. In reality, however, financial data are not stationary. It is possible to significantly over-predict risk for some years, and under-predict it for others, while ending up with a bias statistic very close to 1.

A more relevant question is to study how accurate the risk forecasts were during any 12-month period. For this purpose, we define the rolling 12-month bias statistic for portfolio n,

$$B_n^\tau = \sqrt{\frac{1}{11} \sum_{t=\tau}^{\tau+11} (b_{nt} - \bar{b}_n)^2}, \tag{15.49}$$

where τ denotes the first month of the 12-month window. The 12-month windows are rolled forward 1 month at a time until reaching the end of the observation window. If T is the number of periods in the observation window, then each portfolio will have $T-11$ (overlapping) 12-month windows. It is often informative to consider the frequency distribution of rolling 12-month bias statistics. These histograms quickly indicate whether there were any 12-month periods for which portfolio risk was significantly over-forecast or under-forecast. Figures 15.4 and 15.5 are examples of such distributions. If there are N portfolios, then the histogram will contain $N(T - 11)$ data points.

It is also useful to consider the mean of the rolling 12-month bias statistics, given by

$$\tilde{B} = \frac{1}{N(T-11)} \sum_{n,\tau} B_n^\tau. \tag{15.50}$$

This number indicates whether, on average, risk was over-forecast or under-forecast during the observation window.

Again, it is possible to over-forecast the risk of some portfolios within certain time periods, and to under-forecast the risk of other portfolios during different time periods, while producing a mean rolling bias statistic $\tilde{B}$ close to 1. What is needed is a way to penalize every deviation away from the ideal bias statistic of 1. The 12-month *rolling absolute deviation* (RAD) for portfolio n, defined as

$$\mathrm{RAD}_n = \frac{1}{T-11} \sum_{\tau=1}^{T-11} |B_n^\tau - 1|, \tag{15.51}$$

captures this effect. Smaller RAD numbers, of course, are preferable to larger ones. To assess whether an observed RAD is statistically significant or not, we must consider the properties of the RAD distribution.

In Fig. 15.7, we plot the probability distributions for 12-month RAD over observation windows of 12, 36, and 150 months. Results were obtained via numerical simulation using normally distributed returns. The plot labeled by $T = 12$ corresponds to a single period. The most likely RAD occurs near zero in this case, and the mean of the distribution is approximately $\sqrt{1/12\pi}$, or 0.163. A more precise computation, which accounts for small sample sizes, gives a mean of 0.17.

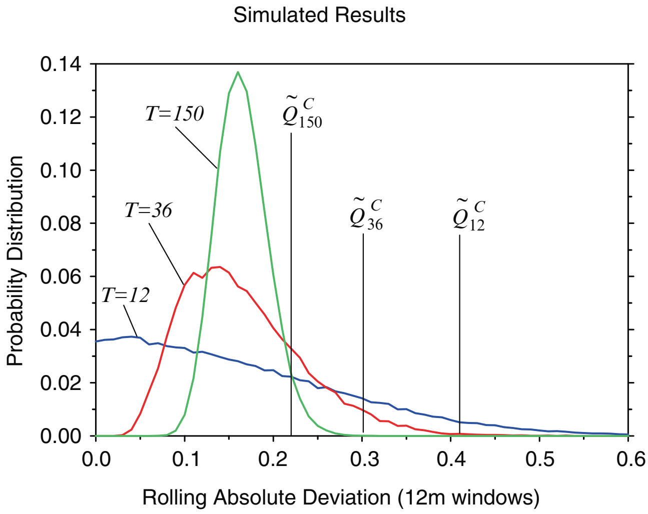


Fig. 15.7 Probability distribution of 12-month rolling absolute deviation (RAD). Results were simulated using normally distributed returns, for *observation* windows of 12, 36, and 150 months. Since a 12-month rolling window within a 12-month observation window corresponds to a single fixed window, the curve labeled $T = 12$ is peaked near zero. A 36-month observation window, however, corresponds to roughly three independent observations, and the peak of the distribution shifts to the right. As the number of periods T becomes large, the central limit theorem assures that the distribution converges to a normal distribution centered at approximately 0.17. Critical values $\tilde{Q}_T^C$ are indicated by the vertical lines. They indicate the 95th percentile of the cumulative distribution

An observation window of $T = 36$ with rolling 12-month windows corresponds to roughly three independent (i.e., non-overlapping) observations. The peak of the probability distribution shifts to the right, and the probability of observing an RAD near zero practically vanishes. This makes intuitive sense, since an RAD of zero would require bias statistics of 1 during *every* 12-month sub-period.

Increasing the observation window to 150 months leads to more than 12 completely independent observations. In this case, the distribution becomes narrower, and the peak shifts further to the right. As T increases further still, the central limit theorem assures that the distribution becomes more sharply peaked about the mean value 0.17, with an approximately normal distribution.

In Fig. 15.7, we indicate by vertical lines the critical mean RAD values, denoted $\tilde{Q}_T^C$, for rolling 12-month windows. These represent the 95th percentile of the cumulative RAD distributions. In other words, assuming perfect risk forecasts and normally distributed returns, over a 36-month observation window we would expect to find an RAD above 0.30 only 5% of the time. If we observe an RAD above this value, we reject the null hypothesis that the risk forecast was accurate.

In reality, of course, returns are not normally distributed. In Fig. 15.8, we plot the critical RAD versus size of observation window T, for various levels of kurtosis. Not surprisingly, the normal distribution ($k = 3$) has the smallest critical values. One useful case to consider is the $T \to \infty$ limit, which converges to a critical value of 0.17 for kurtosis $k = 3$. As kurtosis increases, however, the critical values also increase. For instance, using the same 36-month observation window as before but now assuming a kurtosis $k = 5$, RAD values in excess of 0.38 would occur about 5% of the time.

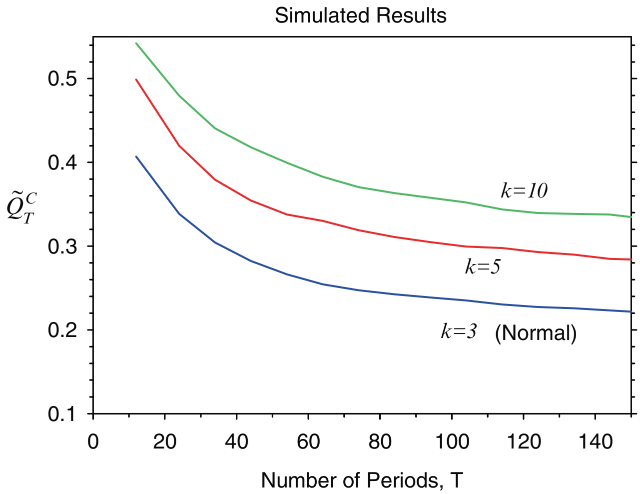


Fig. 15.8 Critical values of rolling absolute deviation (RAD), for rolling 12-month windows, versus the number of periods in the observation window. For perfect risk forecasts and normally distributed returns, over a 60-period observation window we expect to observe an RAD above 0.26 only 5% of the time. For a kurtosis of 10, however, the critical value at 60 periods rises to about 0.40

References

Dempster, A.P., Laird, N.M., and D.B. Rubin, 1977. “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *Journal of the Royal Statistical Society*, vol. 39, no. 1: 1–38.

Grinold, R., Rudd, A., and D. Stefek, Fall 1989. “Global Factors: Fact or Fiction,” *Journal of Portfolio Management*, vol. 16, no. 1: 79–88.

Heston, S.L., and K.G. Rouwenhorst, 1994. “Does Industrial Structure Explain the Benefits of International Diversification?” *Journal of Financial Economics*, vol. 36: 3–27.

Markowitz, H.M., March 1952. “Portfolio Selection,” *Journal of Finance*, vol. 7, no. 1: 77–91.

MSCI Barra Fundamental Data Methodology Handbook, 2008, http://www.msci.com/methodology/ meth_docs/MSCI_Sep08_Fundamental_Data.pdf

Newey, W., and K. West, 1987. “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix,” *Econometrica*, vol. 55, no. 3: 703–708.

Rosenberg, B., March 1974. “Extra-Market Components of Covariance Among Security Prices,” *Journal of Financial and Quantitative Analysis*, vol. 9: 263–274.

Rosenberg, B., and V. Marathe, May 1975. “Tests of Capital Asset Pricing Hypotheses,” *Research Program in Finance*, Working Paper no. 32.

Rudd, A., and B. Rosenberg, 1979. “Realistic Portfolio Optimization,” in *Elton, E., and M. Gruber*, (editors): *Portfolio Theory, 25 Years After, Studies in the Management Sciences*, vol. 11, North Holland, Amsterdam: 21–46.

Sharpe, W.F., September 1964. “Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk,” *Journal of Finance*, vol. 19, no. 3: 425–442.

Tobin, J., February 1958. “Liquidity Preference as a Behavior Toward Risk,” *Review of Economic Studies*, vol. 67: 65–86.

Chapter 16
What Matters Most in Portfolio Construction?
Sensitivity Analysis in the Presence of Transactions Costs

Dean M. Petrich and Ronald N. Kahn

Harry Markowitz invented modern portfolio construction in the 1950s, providing the first precise mathematical approach for trading off expected return against risk. He defined risk as standard deviation and introduced the deceptively simply mean/-variance utility function for portfolio construction. Yet, more than fifty years later, his general framework is still yielding surprising and interesting results.

Much of the recent work on Markowitz portfolio construction has focused on the impact of constraints. Grinold and Kahn (2000a) and Clark et al. (2002) provided detailed estimates of the impact of the long-only constraint. Subsequent work by Clark et al. (2004) expanded the analysis to show the benefit of partial short (e.g., 130/30) strategies over pure long-only strategies. Investigating a different constraint, Johnson et al. (2007) estimated the impact of leverage constraints on performance.

This paper moves from constraints to another complication of "real world" portfolio construction: transactions costs. We know that transactions costs erode performance. But, given the presence of transactions costs, what are the impacts of increasing return forecasting skill, lowering costs, or increasing cost forecasting skill?

This work recalls earlier work by Markowitz (2003), Michaud (1989), and Jorion (1996) on the impact on portfolio construction of errors in estimating risk. In this work, however, we take risk as estimated with perfect accuracy, and focus on the other inputs to portfolio construction: expected returns and transactions costs.

As we move from the early work of Markowitz to the current real world portfolio construction, we tend to move beyond analytic results. In this paper, in particular, we will present the results of simulations. We will take the dynamics of our return forecasts and the structure of transactions costs as given and simulate portfolios subject to different alpha strengths and transaction cost levels. To make the study as transparent as possible, we will evaluate the results strictly on an ex ante basis, and moreover, limit the simulations to a single asset. The one-asset case is also the multiasset case, under the additional assumptions that all asset returns and costs

D.M. Petrich
Head of Proprietary Analytics Group, Barclays Global Investors, San Francisco, CA, USA

R.N. Kahn
Global Head of Equity Research at Barclays Global Investors, San Francisco, CA, USA

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_16,
© Springer Science+Business Media, LLC 2010

are uncorrelated. In the more general case of correlated assets, we expect our basic results to still hold.

Finally, we present some simple observations on the subject of optimization in the presence of errors.

Beyond innate interest, this study has very practical implications. Investment managers allocate research resources to return and cost forecasting and to lowering overall costs. This study provides quantitative estimates of the value of those separate activities. Given, then, estimates of the costs of these activities, investment managers can optimally allocate their research resources.

16.1 Framework

We focus on the Markowitz mean/variance framework applied to active management, as described in Grinold and Kahn (2000b):

$$U = \alpha h - \lambda \omega^2 h^2. \tag{16.1}$$

Here, α is the forecast active return (relative to a benchmark, which can be cash), ω is the active risk, and h is the active holding (the holding relative to the benchmark holding). For our purposes here, we will limit our analysis to the case of one stock. The risk aversion parameter, λ, captures individual investor preference. Introducing the *information ratio* (*IR*):

$$\mathrm{IR} \equiv \frac{\alpha}{\omega}, \tag{16.2}$$

we can show that at optimality:

$$\mathrm{Max}\{U\} = \frac{\mathrm{IR}^2}{4\lambda}. \tag{16.3}$$

According to (16.3), while investors may differ in their risk aversion, they all agree that higher information ratios lead to higher utility. We will use the information ratio as our figure of merit. As we change, e.g., return forecasting skill, what is the impact on information ratio?

Now let us introduce transactions costs into the Markowitz utility function, (16.1). One challenge here is that the return and risk components of the utility function are flow variables – they describe return and risk experienced over time. In contrast, the transactions costs occur at an instant in time. To convert these costs to expected costs over time, we amortize these costs over some average horizon. Mathematically, this will introduce a scaling constant adjustment to the estimated costs:

$$U = \alpha h - \lambda \omega^2 h^2 - \tau f\{h - h_0\}. \tag{16.4}$$

The constant τ is the *transaction cost amortization factor* and $f\{h - h_0\}$ is the estimated cost of trading from an initial position, h_0, to new position, h. Based on our discussion, the transaction cost amortization factor should roughly correspond to the inverse of our expected trading horizon.

We will later find it useful to introduce two trading variables:

$$h \equiv h_0 + p - s, \tag{16.5}$$

where p measures purchases and s measures sales, and we define both as nonnegative numbers. Given that the utility is convex, in any one period, at least one of these two variables will be zero.

16.2 The Detailed Model

Equation (16.4) describes our utility function. Remember that our goal is to understand the impacts of increasing return forecasting skill, lowering costs, or increasing cost forecasting skill, in the presence of transactions costs. To do that, we must go from the general utility function to detailed models of the dynamics of our return forecasts and our transactions cost forecasts.

Let us start with the return forecasts. We assume the forecast alpha in (16.4) follows a mean-reverting process:

$$\alpha(t) = \gamma\alpha(t - \Delta t) + \sigma_\alpha \sqrt{1 - \gamma^2}\varepsilon_\alpha(t), \tag{16.6}$$

with $0 \leq \gamma < 1$. So, our forecast for time t depends in part on the prior forecast (at time $t - \Delta t$) and in part on uncorrelated new information driven by the standardized (mean 0, standard deviation 1) random variable, $\varepsilon_\alpha(t)$. And to be explicit, we assume that the forecast alpha is an annualized quantity, so that $\alpha\Delta t$ is the forecast active return over time span Δt.

We can use (16.6) to specify several properties of the forecast alphas. The mean forecast alpha is zero, and the variance of the forecast alpha is:

$$\text{Var}\{\alpha\} = E\{\alpha^2\} = \sigma_\alpha^2. \tag{16.7}$$

Separately, we know that forecast alphas have the form:

$$\alpha = \text{IC}\omega z, \tag{16.8}$$

where IC is the information coefficient (IC, the correlation of forecasts and subsequent realized active returns) and z is a score, with mean 0 and standard deviation 1 over time. The information coefficient is a critical player in this analysis, as we ultimately want to know the impact of changing it, i.e., changing our return forecasting skill, in the presence of transactions costs.

In general, as follows from (16.8):

$$\text{Var}\{\alpha\} = \text{IC}^2\omega^2. \tag{16.9}$$

Combining (16.7) and (16.9):

$$\sigma_\alpha = \mathrm{IC}\omega. \tag{16.10}$$

What about the half life of the alpha? Since the random component in (16.6) has mean zero, we can compute the expected value of an alpha in the future based on the alpha value today. We recursively calculate the expected alpha J steps in the future ($t = J\Delta t$), based on the alpha value today ($t = 0$) as follows:

$$\begin{aligned} &E\{\alpha(t)\} = \gamma\alpha(t - \Delta t) = \gamma^2\alpha(t - 2\Delta t) = \ldots \\ &\Rightarrow E\{\alpha(t)\} = \gamma^J \alpha(0). \end{aligned} \tag{16.11}$$

We define the half life as the period at which the expected alpha has dropped to half its original value. Using (16.11) and $t = J\Delta t$, the equation becomes:

$$t_{1/2} = \frac{-\ln 2\Delta t}{\ln \gamma}. \tag{16.12}$$

Next, let us tackle the details of our model of transactions costs. We will assume the expected transactions costs have the form:

$$\text{cost} = \left(\text{es} + \text{mi}\sqrt{\frac{p+s}{\text{BT}}}\right)(p+s). \tag{16.13}$$

The cost of a trade has two components. The term linear in purchases plus sales captures bid/ask spread plus any commission costs. The nonlinear term captures the market impact of the trade, and for convenience, we have defined BT as the ratio of average daily volume to fund size.[1] As can be seen from (16.13), we assume that the price impact is proportional to the square root of the trade size and then that the return impact is proportional to the price impact times the amount of the trade. We will let es and mi fluctuate as functions of time in a structured way.

To clarify the meaning of the transactions costs in (16.13), it is helpful to consider an analogy with alphas. The unconditional expected return is zero, and we define the alpha as the expected return conditioned on our information at that time. Alphas fluctuate over time because the information fluctuates over time. The size of the fluctuations, given by (16.9), depends on forecasting skill through the IC, and the bigger the better. By analogy, there is an unconditional transaction cost forecast (which is, however, nonzero). There is also a *conditional* expected transaction cost, which fluctuates over time because our information about transactions costs fluctuates over time. It is the conditional transactions costs expectations that we use in portfolio optimization. The transactions costs in (16.13) are the conditional

[1] Hence the term inside the square root in (16.13) represents the trade as a fraction of average daily volume. This is consistent with modeling market impact as proportional to daily stock volatility times the square root of trade size as a fraction of average daily volume. See Grinold and Kahn (2000b) for more details.

transactions cost forecasts. By analogy with the alpha IC, the larger the fluctuations in the conditional cost forecasts, the greater the explanatory power of our transactions costs forecasts. In the framework of this note, that leads unambiguously to improved portfolio performance because we're able to trade opportunistically, when the transactions costs are low.

Notice that although the transactions costs fluctuate, the fluctuations are not "errors." By analogy with the alphas again, if we optimize with an alpha of 2% and the conditional expected return is 2%, there is no error, even if the ex post realized return is something other than 2%. There is, however, an error if we optimize with an alpha of 2%, and for some unknown reason, the conditional expected return is, say, 3%. Applying this logic to transactions costs, if we optimize with a transactions cost forecast of 25 basis points (bps) and the conditional expected transactions cost is 25 bps, there is no error even if the realized transactions cost is 50 bps. There is an error if we optimize with 25 bps and, for some unknown reason, the expected cost is, say, 40 bps. In this note, we are concerned primarily with the error-free case where our forecasts exactly match the mean of the ex post distribution, although we have one short section toward the end that addresses errors.

We want to be able to separately calibrate the gains associated with

1. Lowering the overall trading costs (the unconditional transactions cost) and
2. Improving transactions costs forecasting skill (the size of the fluctuations around the unconditional cost, generated by our use of information at the time).

Consistent with the discussion in the previous two paragraphs, we will use a model for es and mi that lets us control both the unconditional and conditional transactions costs expectations. We assume the transactions cost coefficients of (16.13) are drawn from a lognormal distribution. Each period has an independent draw:

$$\begin{aligned} \mathrm{es}(t) &= \exp\{m_{\mathrm{es}} + s_{\mathrm{es}}\varepsilon_{\mathrm{es}}(t)\} \\ \mathrm{mi}(t) &= \exp\{m_{\mathrm{mi}} + s_{\mathrm{mi}}\varepsilon_{\mathrm{mi}}(t)\}, \end{aligned} \tag{16.14}$$

where $\varepsilon_{\mathrm{es}}$ and $\varepsilon_{\mathrm{mi}}$ are mean 0, standard deviation 1 random variables.

Once again, for emphasis, note that (16.14) does not represent the ex post cost distribution, and moreover the fluctuations created by $\varepsilon_{\mathrm{es}}$ and $\varepsilon_{\mathrm{mi}}$ are not related to errors. The terms $\mathrm{es}(t)$ and $\mathrm{mi}(t)$ are perfect ex ante forecasts for time t. The mean of the ex post distribution for es at time t is exactly $\mathrm{es}(t)$, and the mean of the ex post distribution for mi at time t is exactly $\mathrm{mi}(t)$. The transactions costs forecasts fluctuate because our information about the market fluctuates.

Starting from (16.14), the mean, μ, and standard deviation, σ, of these estimates are:

$$\begin{aligned} \mu_{\mathrm{es}} &= \exp\left\{m_{\mathrm{es}} + \tfrac{1}{2}s_{\mathrm{es}}^2\right\}, \\ \frac{\sigma_{\mathrm{es}}^2}{\mu_{\mathrm{es}}^2} &= \exp\left\{s_{\mathrm{es}}^2\right\} - 1. \end{aligned} \tag{16.15}$$

with similar expressions for the market impact parameter estimates.

As discussed above, the parameters s_{es} and s_{mi} represent the power of the cost model to explain fluctuations about the mean value. We take the mean realized cost as given, so that these are related to the R^2 of the cost model. Similarly, the parameters μ_{es} and μ_{mi} are akin to mean realized costs.

16.3 Asset Position Dynamics

Now that we have specified the dynamics of our return forecasts and the form of our forecast and the realized transactions costs, we can maximize the utility function and solve for the dynamics of our asset position.

Assume we are purchasing. Then the optimal purchase p satisfies:

$$\alpha - 2\lambda\omega^2(h_0 + p) - \tau \cdot \left(\mathrm{es} + \frac{3}{2}\mathrm{mi}\sqrt{\frac{p}{\mathrm{BT}}}\right) = 0. \tag{16.16}$$

Equation (16.16) is a quadratic equation for $\sqrt{p}$:

$$p + \frac{3\mathrm{mi}\tau}{4\sqrt{\mathrm{BT}}\lambda\omega^2}\sqrt{p} - \frac{(\alpha - 2\lambda\omega^2 h_0 - \tau\mathrm{es})}{2\lambda\omega^2} = 0. \tag{16.17}$$

Define:

$$\begin{aligned} \mathrm{mu} &= \alpha - 2\lambda\omega^2 h_0 \\ \phi &= \tfrac{3\tau}{8\lambda\omega^2\sqrt{\mathrm{BT}}}, \end{aligned} \tag{16.18}$$

then:

$$p + 2\phi\mathrm{mi}\sqrt{p} - \frac{(mu - \tau\mathrm{es})}{2\lambda\omega^2} = 0, \tag{16.19}$$

$$\sqrt{p} = -\phi\mathrm{mi} + \sqrt{\phi^2\mathrm{mi}^2 + \left(\frac{\mathrm{mu} - \tau\mathrm{es}}{2\lambda\omega^2}\right)}. \tag{16.20}$$

We will only purchase if mu $>$ τes. Similarly, for sales, we find that:

$$\sqrt{s} = -\phi\mathrm{mi} + \sqrt{\phi^2\mathrm{mi}^2 + \left(\frac{-\mathrm{mu} - \tau\mathrm{es}}{2\lambda\omega^2}\right)}. \tag{16.21}$$

We will only sell if mu $< -\tau$es. Hence, we find a no-trade zone when $-\tau$es $<$ mu $< \tau$es.

16.4 Simulations

For a given set of parameters, we simulated results – including the expected alpha, risk, turnover, and transactions costs – over different draws of alphas and costs. For each combination of input parameters, we ran a two-million period simulation, corresponding to a 40,000 year simulation, since we fixed Δt, the time between rebalances, to 1 week. To make comparisons of different parameter values more meaningful, all the simulations shared the same time series of noise terms.

In addition to Δt, we fixed several parameters as constant over all the simulations:

- $\sigma_\alpha \approx 1.8\%$.
- $\omega = 30\%$.
- The alpha half life = 3 months, or $\gamma = 0.95$.
- BT = 0.1, meaning that a 0.1% trade generates market impact, mi.

With these all fixed, we ran simulations that varied in the mean and standard deviation of the transaction cost coefficients:

- μ_{es}, 5–50 basis points, step size five basis points
- σ_{es}, 0–40 basis points, step size five basis points
- μ_{mi}, 50–95 basis points, step size five basis points
- σ_{mi}, 0–40 basis points, step size five basis points

For each combination of the above, we adjusted the risk aversion to achieve a forecast risk of six basis points ± 0.006 basis points. From where does the six basis point figure come? Assume that assets have 30% risk, and that a 1,000 asset portfolio has 2% risk. If all assets are uncorrelated, each asset accounts for six basis points of risk in the portfolio.

Now, what about the transaction cost amortization factor? In principle, given the alpha dynamics and the transactions costs, there should be one choice of this factor that will maximize performance. To facilitate this, we simply ran simulations for $\tau = 1, 2, \ldots, 15$. Then, for any combination of other variables, we chose the τ simulation with the highest realized IR. For some parameters, it was impossible to achieve the target risk, given our range of τ. In that case, we dropped the observation.

What about the information coefficient? That should have the largest impact on performance, and yet we appear not to vary it in these simulations. In fact, we effectively varied the information coefficient by postprocessing the results.

The mean-variance utility function does not depend on the overall scale, so we can derive results for all IC levels, given results at one IC level. For instance, say, we run a simulation with a given σ_α, λ, and τ. We would get the same history of portfolios if we ran a simulation with $\sigma'_\alpha = \kappa\sigma_\alpha$, $\lambda' = \kappa\lambda$, and $\tau' = \kappa\tau$, where κ is any positive constant.[2] This history of portfolios will have the same risk and cost levels (because the positions do not change), but the expected alpha will differ by a factor of κ. Therefore, we need not re-run the simulation for σ'_α, λ', and τ'; we need to only re-interpret the existing results for σ_α, λ, and τ. This results in an enormous time savings. We use this trick to derive results for five different IC values (from 0.03 to 0.30), with no additional simulations.

[2] The trade size and no-trade region are unchanged under the above transformation.

16.5 Simulation Results

The above specification corresponds to a large number of large simulations. Ignoring the simulations over different values of τ, we have over 40,000 different parameter combinations, each one with a corresponding two-million period simulation. Needless to say, we will not present all those results here.

Instead, we will show a few representative numbers and then describe how we have summarized all the results. Table 16.1 provides the representative numbers.

The last column in Table 16.1 lists transfer coefficients for these simulations. The transfer coefficient is the ratio of the realized IR to the intrinsic (no transaction cost) IR. Note that in this case of just one asset, the intrinsic IR equals the IC.

To summarize all the data, we focused on answering the question, if we move a parameter away from the in-sample mean, how much does the IR change? This is a sensitivity analysis, which we quantified using the following regression:

$$\frac{\text{IR}}{\langle\text{IR}\rangle} = \beta_0+\beta_1\frac{\mu_{\text{es}}}{\langle\mu_{\text{es}}\rangle}+\beta_2\frac{\mu_{\text{mi}}}{\langle\mu_{\text{mi}}\rangle}+\beta_3\frac{\sigma_{\text{es}}}{\langle\sigma_{\text{es}}\rangle}+\beta_4\frac{\sigma_{\text{mi}}}{\langle\sigma_{\text{mi}}\rangle}+\beta_5\frac{\sigma_{\alpha}}{\langle\sigma_{\alpha}\rangle}+\text{noise}. \quad (16.22)$$

In (16.22), the $\langle\rangle$ operator represents an average over all 40,000 + different cost and alpha scale combinations in the data set. To interpret the coefficients, note that an $x\%$ change in the alpha scale, say, will lead to a $\beta_5 \cdot x\%$ change in the IR. Table 16.2 displays the results.

Table 16.1 Representative simulations

μ-es (%)	μ-mi (%)	σ-es (%)	σ-mi (%)	std(α) (%)	IC	IR	TC
0.25	0.75	0.2	0.15	3	0.10	0.043	0.43
0.25	0.75	0.2	0.2	3	0.10	0.044	0.44
0.25	0.75	0.2	0.25	3	0.10	0.044	0.44
0.25	0.75	0.2	0.15	5	0.17	0.092	0.54
0.25	0.75	0.2	0.2	5	0.17	0.093	0.54
0.25	0.75	0.2	0.25	5	0.17	0.093	0.55
0.25	0.75	0.2	0.15	7	0.23	0.147	0.64
0.25	0.75	0.2	0.2	7	0.23	0.148	0.64
0.25	0.75	0.2	0.25	7	0.23	0.149	0.65

Table 16.2 Sensitivity analysis

	In-sample mean (%)	Coefficient	t-stat
Intercept		0.1156	54
μ-es	0.275	−0.1714	−257
μ-mi	0.725	−0.2810	−160
σ-es	0.2	0.0356	66
σ-mi	0.2	0.0344	64
$\sigma - \alpha$	5	1.2667	2,058

For example, in this data set, a 10% reduction in the market impact coefficient leads to a 2.8% increase in IR, and a 10% increase in σ_{es} leads to a 0.36% increase in IR. The coefficient on alpha scale exceeds one, therefore IC increases result in even larger IR increases: a 10% increase in IC results in a 12.7% increase in IR. The Appendix contains a simple argument why this coefficient must exceed one.

The coefficients indicate that for this range of parameters increasing IC is the most valuable by far, followed by

- Decreasing the average market impact coefficient
- Decreasing the average fixed cost
- Improving forecasts of fixed costs around its mean
- Improving forecasts of market impact around its mean

Again, note that these results are sensitive to the in-sample means. If we ran a different set of simulations, say with higher transactions costs and lower IC, the sensitivity analysis would yield somewhat different results.

16.6 Errors in Parameter Values

The analysis so far has assumed we know all parameters perfectly, in advance. However, this is implausible; for example, we may optimize thinking σ_α=5%, but in fact, our IC is weaker, and in truth, σ_α=4%. What is the loss associated with that?

Say that the alpha scale we use for optimization is off by a factor of κ:

$$\alpha_{\text{true}} = \kappa\alpha_{\text{opt}}. \tag{16.23}$$

Then, the IRs are related by:

$$\text{IR}_{\text{true}} = \frac{\alpha_{\text{true}} - \text{TC}}{\text{risk}} = \frac{\kappa\alpha_{\text{opt}} - \text{TC}}{\text{risk}} = (\kappa - 1)\frac{\alpha_{\text{opt}}}{\text{risk}} + \text{IR}_{\text{opt}}. \tag{16.24}$$

Using the numbers above, say, we optimize thinking $\sigma_\alpha = 5\%$, but in fact our IC is such that $\sigma_\alpha = 4\%$. For transaction costs, take representative numbers es = 25 basis points and mi = 75 basis points, with no fluctuations. Picking this row out of the data set, we find an optimal alpha of 0.81%, costs of 0.28%, risk of six basis points, and optimal IR of 0.088. Therefore:

$$\text{IR}_{\text{true}} = \left(\frac{4}{5} - 1\right)\frac{0.81}{6} + 0.088 = 0.061. \tag{16.25}$$

We thought we were going to get IR = 0.088, but in fact, we realized only IR = 0.061. Note that based on our presumed values, our IC error was 20%, but it resulted in an IR error of over 30%!

There are two sources of loss here. First, there is the obvious loss that we are getting less alpha than we thought. A secondary loss is that we believe our alpha is stronger than it is, so we end up trading more than we would if we knew the true alpha scale.[3]

We can similarly calculate the impact of errors in cost scale, but because the alpha is larger than the costs (in this example, by a factor of 81/28), errors in alpha scale matter more than errors in costs. However, clearly, the lower the information ratio, the larger the adverse impact of any transaction cost scale errors.

16.7 Summary

Markowitz optimization is a standard approach for portfolio construction. Within that framework, we have explored the impact of increasing return forecasting skill, lowering costs, and increasing cost forecasting skill in the presence of transactions costs. In particular, we have run an extensive set of simulations to probe the sensitivity of realized information ratios to changes in information coefficient, average transactions costs, and our ability to forecast transactions costs.

In summary, we find that the information ratio is most sensitive to, in order,

- The information coefficient
- Mean transactions costs
- Our ability to forecast fluctuations about the mean transactions cost.

Each item is about five times more relevant than the succeeding item.

This paper quantifies the benefit of improving these parameters, but says nothing about the costs to do so.

Acknowledgment The authors would like to acknowledge Mark Coppejans, Ananth Madhavan, and particularly, Richard Grinold for helpful discussions.

Appendix: Why Increases in IC Lead to Even Greater Increases in IR

This appendix indicates why an $x\%$ increase in IC leads to an even larger percentage increase in IR. It is essentially the same calculation as in the "Errors in Parameter Values" section, viewed in a different way.

[3] This is why it's not correct to compare the $\sigma_\alpha = 5\%$ and $\sigma_\alpha = 4\%$ observations in the data set to gauge the effect of errors. The observations in the data set have the alpha, costs and IR that apply if the parameters are known perfectly. Different levels of σ_α have different turnover levels. In the case where we think $\sigma_\alpha = 5\%$, but in truth $\sigma_\alpha = 4\%$, we would run our portfolio with too high a turnover level. Therefore the *IR* will be lower than if we knew $\sigma_\alpha = 4\%$ in advance.

Say, we start with an alpha scale of σ_α. Running this strategy yields a precost alpha of α and transaction costs TC. Now, imagine we discover some new signals that increase our IC, and hence increase σ_α to $\kappa\sigma_\alpha$ (where $\kappa > 1$). However, we continue to rebalance as if σ_α was unchanged. Our trading rule is the same, but because our signals are better, our portfolio alpha is also better; in particular, over time, our portfolio will have a precost alpha of exactly $\kappa\alpha$.

In terms of IR:

$$\begin{aligned} \mathrm{IR}_{\mathrm{new}} &= \tfrac{\kappa\alpha - \mathrm{TC}}{\mathrm{risk}} = (\kappa - 1)\tfrac{\alpha}{\mathrm{risk}} + \mathrm{IR}_{\mathrm{old}} \\ &\Rightarrow \left(\tfrac{\Delta\sigma_\alpha}{\sigma_\alpha}\right)\tfrac{\alpha}{\mathrm{risk}} + \mathrm{IR}_{\mathrm{old}}. \end{aligned} \tag{16.26}$$

Put another way:

$$\Delta\mathrm{IR} = \frac{\Delta\sigma_\alpha}{\sigma_\alpha}\frac{\alpha}{\mathrm{risk}}. \tag{16.27}$$

Recall that here α is the precost alpha, so we have $\frac{\alpha}{\mathrm{risk}} \geq \mathrm{IR}$. Therefore:

$$\frac{\Delta\mathrm{IR}}{\mathrm{IR}} = \frac{\Delta\sigma_\alpha}{\sigma_\alpha}\frac{\alpha}{\mathrm{risk}}\frac{1}{\mathrm{IR}} \geqslant \frac{\Delta\sigma_\alpha}{\sigma_\alpha}. \tag{16.28}$$

This argument guarantees that the IR gain will exceed the IC gain, even at a fixed turnover level. If we allow ourselves to tune the turnover to reflect the new level of IC, we will do even better.

A few things to note:

- This effect is due to transaction costs and holds for multiasset portfolios, not only this one-asset example.
- The larger the transaction costs, the more pronounced this effect.
- The transfer coefficient is an increasing function of IC. One way to look at this is that the larger the IC, the less the transaction costs matter and the closer the portfolio gets to the ideal transfer coefficient of 1.

References

Clark, Roger; de Silva, Harindra; and Thorley, Stephen. "Portfolio Constraints and the Fundamental Law of Active Management." *Financial Analysts Journal*, Vol. 58, No. 5 (September/October 2002), pp. 48–66.

Clark, Roger; de Silva, Harindra; and Sapra, Steven. "Toward More Information-Efficient Portfolios." *Journal of Portfolio Management*, Fall 2004, pp. 54–62.

Fabozzi, Frank J.; Kolm, Petter N.; Pachamanova, Dessislava A.; and Focardi, Sergio M. *Robust Portfolio Optimization and Management*. (Hoboken, NJ: Wiley Finance, 2007).

Grinold, Richard C., and Kahn, Ronald N. *Active Portfolio Management*, 2nd Edition. (New York: McGraw-Hill, 2000a).

Grinold, Richard C., and Kahn, Ronald N. "The Efficiency Gains of Long-Short Investing." *Financial Analysts Journal*, Vol. 56, No. 6 (November/December 2000b).

Johnson, Seanna; Kahn, Ronald N.; and Petrich, Dean. "Optimal Gearing." *Journal of Portfolio Management*, Summer 2007, pp. 10–18.

Jorion, Philippe. "R^2: Measuring the Risk in Value-At-Risk." *Financial Analysts Journal*, Vol. 52, No. 6 (November/December 1996).

Markowitz, Horry, and Usmen, Nilufer. "Resampled Frontiers vs Diffuse Bayes: An Experiment." *Journal of Investment Management*, Vol. 1, No. 4 (Fourth Quarter 2003).

Michaud, Richard O. "The Markowitz Optimization Enigma: Is 'Optimized' Optimal?" *Financial Analysts Journal*, Vol. 45, No. 1 (January/February 1989).

Michaud, Richard O. *Efficient Asset Management*. (Boston: Harvard Business School Press, 1998).

Chapter 17
Risk Management and Portfolio Optimization for Volatile Markets

Svetlozar T. Rachev, Borjana Racheva-Iotova, Stoyan V. Stoyanov, and Frank J. Fabozzi

17.1 Introduction

The two main conventional approaches for modeling asset returns are based on either a historical or a normal (Gaussian) distribution of returns. Neither approach adequately captures the unusual behavior of asset prices and returns. The historical model is bounded by the extent of the available observations, and the normal distribution model inherently cannot produce extreme returns.

The inadequacy of the normal distribution is well recognized by the risk management community. To quote one major vendor:

"It has often been argued that the true distributions returns (even after standardizing by the volatility) imply a larger probability of extreme returns than that implied from the normal distribution. Although we could try to specify a distribution that fits returns better, it would be a daunting task, especially if we consider that the new distribution would have to provide a good fit across all asset classes." (Technical Manual, RMG, 2001).

There are many studies exploring the nonnormality of assets returns that suggest alternative approaches. Among the well-known candidates are the Student's t distribution, the generalized hyperbolic distributions (see Bibby and Sorensen 2003), and the stable Paretian distributions (see Rachev and Mittnik 2000). At least some of their forms are subordinated normal models and thus provide a very practical and tractable framework. Rachev et al. (2005) provide an introduction to the heavy-tailed models in finance.

S.T. Rachev (✉)
University of Karlsruhe, KIT, University of California at Santa Barbara and FinAnalytica Inc.
e-mail: rachev@kit.edu

B. Racheva-Iotova and S.V. Stoyanov
FinAnalytica Inc.
e-mail: borjana.racheva@finanalytica.com; stoyan.stoyanov@finanalytica.com

F.J. Fabozzi
Yale School of Management
e-mail: frank.fabozzi@yale.edu

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_17,
© Springer Science+Business Media, LLC 2010

In response to these challenges, we use the generalized multivariate stable distributions and generalized risk-factor dependencies, thereby creating a paradigm shift resulting in a consistent and uniform use of the most viable class of non-normal probability models in finance. This chapter discusses the reworking of the classical approaches into a framework that allows for an increased flexibility, accurate assets modeling, and sound risk measurement, employing generalized stable distributions together with the average value-at-risk (AVaR) risk measure, see Rachev et al. (2007).

The chapter is organized as follows. Section 17.1 discusses several heavy-tailed models with special attention to the generalized stable distributions. In Sect. 17.2, we discuss multivariate modeling. Section 17.3 provides a summary of risk and performance measure properties and describes the AVaR measure. Section 17.4 discusses risk budgeting based on AVaR and Sect. 17.5 is devoted to optimal portfolio problems. In Sect. 17.6, we comment on performance measures consistent with AVaR. Section 17.7 contains an empirical example using the Russell 2000 universe.

17.2 Heavy-Tailed and Asymmetric Models for Asset Returns

It is vital to specify the distribution of asset returns for risk management and optimal asset allocation. A failure may lead to significant underestimation of portfolio risk and, consequently, to wrong decisions.

The distributional modeling of financial variables has several dimensions. First, there should be a realistic model for the returns of each financial variable considered separately. That is, we should employ realistic one-dimensional models. Second, the model should capture the dependence between the one-dimensional variables. Therefore, we need a true multivariate model with the above two building blocks correctly specified.

17.2.1 One-Dimensional Models

The cornerstone theories in finance, such as the mean-variance model for portfolio selection and asset pricing models, have been developed based on the assumption that asset returns follow a normal distribution. Yet, there is little, if any, credible empirical evidence that supports this assumption for financial assets traded in most markets throughout the world. Moreover, the evidence is clear that financial return series are heavy-tailed and, possibly, skewed. Fortunately, several studies have analyzed the consequences of relaxing the normality assumption and developed generalizations of prevalent concepts in financial theory that can accommodate heavy-tailed returns (see Rachev and Mittnik (2000) and Rachev (2003) and references therein).

Mandelbrot (1963) strongly rejected normality as a distributional model for asset returns, conjecturing that financial return processes behave like non-Gaussian stable processes. To distinguish between Gaussian and non-Gaussian stable distributions, the latter are commonly referred to as "stable Paretian" distributions or "Levy stable" distributions.[1]

While there have been several studies in the 1960s that have extended Mandelbrot's investigation of financial return processes, the most notable is Fama (1963) and Fama (1965). Fama's and others work led to a consolidation of the stable Paretian hypothesis. In the 1970s, however, closer empirical scrutiny of the "stability" of fitted stable Paretian distributions also produced evidence that was not consistent with the stable Paretian hypothesis. Specifically, it was often reported that fitted characteristic exponents (or tail-indices) did not remain constant under temporal aggregation.[2] Partly in response to these empirical "inconsistencies," various alternatives to the stable law were proposed in the literature, including the fat-tailed distributions in the domain of attraction of a stable Paretian law, finite mixtures of normal distributions, the Student t-distribution, and the hyperbolic distribution, see Bibby and Sorensen (2003).

Recent attacks on Mandelbrot's stable Paretian hypothesis focus on the claim that empirical asset return distributions are not as heavy-tailed as the non-Gaussian stable law suggests. Studies that come to such conclusions are typically based on tail-index estimates obtained with the Hill estimator. Because sample sizes beyond 100,000 are required to obtain reasonably accurate estimates, the Hill estimator is highly unreliable for testing the stable hypothesis. More importantly, Mandelbrot's stable Paretian hypothesis is interpreted too narrowly, if one focuses solely on the *marginal* distribution of return processes. The hypothesis involves more than simply fitting the marginal asset return distributions. Stable Paretian laws describe the fundamental "building blocks" (e.g., innovations) that drive asset return processes. In addition to describing these "building blocks," a complete model should be rich enough to encompass relevant stylized facts, such as

- Non-Gaussian, heavy-tailed, and skewed distributions
- Volatility clustering (ARCH-effects)
- Temporal dependence of the tail behavior
- Short- and long-range dependence

An attractive feature of stable models – not shared by other distributional models – is that they allow us to generalize Gaussian-based financial theories and, thus, to build a coherent and more general framework for financial modeling. The generalizations are only possible because of specific probabilistic properties that are unique to (Gaussian and non-Gaussian) stable laws, namely, the stability property, the Central

[1] Stable Paretian is used to emphasize that the tails of the non-Gaussian stable density have Pareto power-type decay. "Levy stable" is used in recognition of the seminal work of Paul Levy's introduction and characterization of the class of non-Gaussian stable laws.

[2] For a more recent study, see Akgiray and Booth (1988) and Akgiray and Lamoureux (1989).

Limit Theorem, and the invariance principle for stable processes. Detailed accounts of properties of stable distributed random variables can be found in Samorodnitsky and Taqqu (1994) and Janicki and Weron (1994).

Stable distributions are defined by the means of their characteristic functions, $\phi_X(t) = E\,\mathrm{e}^{\mathrm{i}tX}$. The characteristic function has the following form,

$$\phi_X(t) = \begin{cases} \exp\left(-\sigma^\alpha |t|^\alpha \left[1 - \mathrm{i}\beta \frac{t}{|t|} \tan \frac{\pi\alpha}{2}\right] + \mathrm{i}\mu t\right), \alpha \neq 1 \\ \exp\left(-\sigma |t| \left[1 + \mathrm{i}\beta \frac{2}{\pi} \frac{t}{|t|} \log |t|\right] + \mathrm{i}\mu t\right), \alpha = 1 \end{cases}. \tag{17.1}$$

In the general case, no closed-form expressions are known for the probability density and distribution functions of stable distributions. The formula in (17.1) implies that they are described by four parameters: α, called the index of stability, which determines the tail weight or density's kurtosis with $0 < \alpha \leq 2$, β, called the skewness parameter, which determines the density's skewness with $-1 \leq \beta \leq 1$, $\sigma > 0$ which is a scale parameter, and μ which is a location parameter. Stable distributions allow for skewed distributions when $\beta \neq 0$ and when β is zero, the distribution is symmetric around μ. Stable Paretian laws have fat tails, meaning that extreme events have high probability relative to the normal distribution when $\alpha < 2$. The Gaussian distribution is a stable distribution with $\alpha = 2$. (For more details on the properties of stable distributions, see Samorodnitsky and Taqqu 1994.) Of the four parameters, α and β are most important as they identify two fundamental properties that are atypical of the normal distribution – heavy tails and asymmetry.

Rachev et al. (2006) consider the daily return distribution of 382 US stocks in the framework of two probability models – the homoskedastic independent, identical distributed model and the conditional heteroskedastic ARMA-GARCH model. In both models, the Gaussian hypothesis is strongly rejected in favor of the stable Paretian hypothesis which better explains the tails and the central part of the return distribution. The companies in the study were constituents of the S&P 500 with complete history in the 12-year time period from January 1, 1992 to December 12, 2003. Figure 17.1 illustrates the estimated (α, β) pairs from historical data. The estimated parameters suggest significant heavy-tail and asymmetry which are phenomena that cannot be accounted for by the normal distribution.

17.2.2 Multivariate Models

For the purpose of portfolio risk estimation, constructing one-dimensional models for the instruments is incomplete. Failure to account for the dependencies between the instruments may be fatal for the analysis.

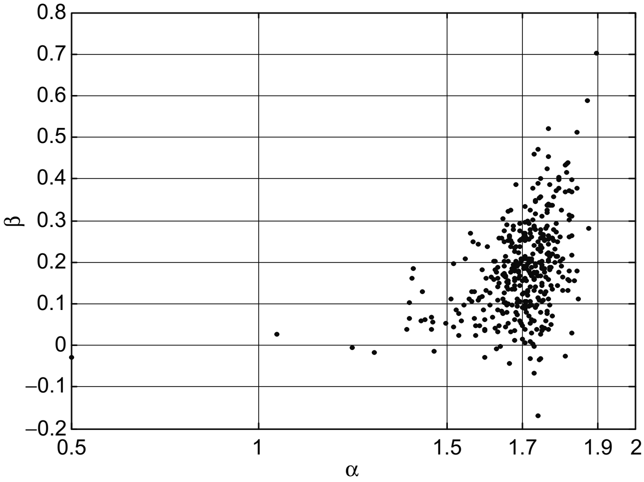


Fig. 17.1 Scatter plot of the stability index and the skewness parameter for the daily returns of 382 stocks (reproduced from Fig. 1.1 in Rachev et al. 2006)

There are two ways to build a complete multivariate model. It is possible to hypothesize a multivariate distribution directly (i.e., the dependence between stock returns as well as their one-dimensional behavior). Assumptions of this type include the multivariate normal, the multivariate Student t, the more general elliptical family, the multivariate stable, etc. Sometimes, in analyzing dependence, an explicit assumption is not made and the covariance matrix is very often relied on. Nevertheless, it should be kept in mind that this is consistent with the multivariate normal hypothesis. More generally, the covariance matrix can describe only linear dependencies and this is a basic limitation.

In the last decade, a second approach has become popular. One can specify separately the one-dimensional hypotheses and the dependence structure through a function called copula. This is a more general and more appealing method because one is free to choose different parametric models for the stand-alone variables and a parametric copula function. For more information, see Embrechts et al. (2003).

17.2.3 Generalized Stable Distribution Modeling

Figure 17.1 indicates that the tail behavior of financial variables may vary. Generalized stable distribution modeling is based on fitting univariate stable distributions for each one-dimensional set of returns or risk factors, each with its own parameter estimates α_i, β_i μ_i, σ_i, $i = 1,2,\ldots,K$, where K is the number of risk factors, along with a dependency structure.

One way to produce the cross-sectional dependency structure is through a scale mixing process (called a "subordinated" process in the mathematical finance literature) as follows.

(a) Compute a robust mean vector and covariance matrix estimate of the risk factors to get rid of the outliers, and have a good covariance matrix estimate for the central bulk of the data.
(b) Multiply each of the random variable component of the scenarios by a strictly positive stable random variable with an index $\alpha_i/2$, $i = 1,2,\ldots,K$. The vector of the stable random variable scale multipliers is usually independent of the normal scenario vectors, but it can also be dependent. See for example Rachev and Mittnik (2000).

Another very promising approach to building the cross-sectional dependence model is through the use of copulas, an approach that is quite attractive because it allows for modeling higher correlations during extreme market movements, thereby accurately reflecting lower portfolio diversification at such times. Section 17.2.4 briefly discusses copulas.

17.2.4 Copula Dependence Models

Correlation is a broad concept in modern finance and insurance, and is a measure of dependence between random variables. However, this term is very often incorrectly used to describe dependence in general. Actually, correlation is one particular measure of dependence among many. In the world of multivariate normal distribution and, more generally in the world of spherical and elliptical distributions, it is an accepted measure.

Financial theories and risk management analysis rely crucially on the dependence structure of asset returns. A major limitation of correlation as a measure of the dependence between two random variables is that zero correlation does not imply independence for non-Gaussian distributions. Furthermore, correlation is symmetric and, in order to be more realistic, we need a more general notion which can reflect the local variation in dependence that is related to the level of returns, in particular, those shapes that correspond to higher correlations with extreme comovements in returns than with small to modest comovements.

From a mathematical viewpoint, a copula function C is nothing more than a probability distribution function on the n-dimensional hypercube

$$C(u_1, u_2, \ldots, u_n), \quad u_i \in [0, 1] \text{ for } i = 1, 2, \ldots, n,$$

where $C(u_i) = u_i, i = 1, \ldots n$.

It is known that for any multivariate cumulative distribution function:

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \le x_1, X_2 \le x_2, \ldots X_n \le x_n),$$

there exists a copula C such that

$$F(x_1, x_2, \ldots, x_n) = C(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)),$$

where the $F_i(x_i)$ are the marginal distributions of $F(x_1, x_2, \ldots, x_n)$, and conversely for any copula C the right-hand side of the above equation defines a multivariate distribution function $F(x_1, x_2, \ldots, x_n)$. See, for example, Bradley and Taqqu (2003), Sklar (1996), and Embrechts et al. (2003).

The main idea behind the use of copulas is that one can first specify the marginal distributions in a sensible manner (e.g., fitting the marginal distribution models to risk factor data, and then specifying a copula C to capture the multivariate dependency structure in the best suited manner).

A possible approach for choosing a flexible copula model is to adopt the copula of a parametric multivariate distribution. In this way, the copula itself will have a parametric form. There are many multivariate laws mentioned in the literature, which can be used for this purpose. One such example is the Gaussian copula, i.e., the copula of the multivariate normal distribution. It is easy to work with, but has one major drawback: It implies that extreme events are asymptotically independent. Thus, the probability of joint occurrence of large negative returns of two stocks is significantly underestimated. An alternative to the Gaussian copula is the Student's t copula (i.e., the copula of the multivariate Student's t distribution). It better models the probability of joint extreme events but has the disadvantage of being symmetric. Thus, the probability of joint occurrence of very large returns is the same as the probability of joint occurrence of very small returns. This deficiency is not present in the skewed Student's t copula which we believe is a much more realistic model of dependency. This is the copula of the multivariate skewed Student's t distribution defined by the following stochastic representation,

$$X = \mu + \gamma W + Z\sqrt{W},$$

where $W \in \mathrm{IG}(\nu/2, \nu/2)$, i.e., W is inverse gamma distributed, Z is the multivariate normal random variable, $Z \in N_n(0, \Sigma)$, W and Z are independent, and the constants μ and γ are such that the sign of a given component of γ controls the asymmetry of the corresponding component of X, and μ is a location parameter contributing to the mean of X. The skewed Student's t copula has the following parametric form,

$$C(u_1, \ldots, u_n) = \int_{-\infty}^{t_{\nu,\gamma}^{-1}(u_1)} \cdots \int_{-\infty}^{t_{\nu,\gamma}^{-1}(u_n)} f(x_1, \ldots, x_n)\mathrm{d}x_1 \ldots \mathrm{d}x_n,$$

where $t_{\nu,\gamma}^{-1}(u_1)$ is the inverse cdf of the one-dimensional skewed Student's t distribution and $f(x_1, \ldots, x_n)$ is the density of the multivariate skewed Student's t distribution,

$$f(x_1,\dots,x_n) = \frac{2^{1-(\nu+n)/2}}{\Gamma(\nu/2)(\pi\nu)^{n/2}\,|\Sigma|^{1/2}} * \frac{\exp\left((x-\mu)'\Sigma^{-1}\gamma\right)}{\left(1+\dfrac{(x-\mu)'\Sigma^{-1}(x-\mu)}{\nu}\right)^{\frac{\nu+n}{2}}} \\ * \frac{K_{\frac{\nu+n}{2}}\left(\sqrt{(\nu+(x-\mu)'\Sigma^{-1}(x-\mu))\,\gamma'\Sigma^{-1}\gamma}\right)}{\left(\sqrt{(\nu+(x-\mu)'\Sigma^{-1}(x-\mu))\,\gamma'\Sigma^{-1}\gamma}\right)^{-\frac{\nu+n}{2}}},$$

where $K_\lambda(x)$ stands for the modified Bessel function of the third kind. The skewed Student's t copula has the following features which make it a flexible and attractive model:

- It has a parametric form which makes the copula an attractive model in higher dimensions.
- The underlying stochastic representation facilitates scenario generation from the copula.
- It can describe the tail dependence if present in the data.
- It can describe the asymmetric dependence if present in the data.

17.3 Average Value-at-Risk

A major activity in many financial institutions is to recognize the sources of risk, then manage and control them. This is possible only if risk is quantified. If we can measure the risk of a portfolio, then we can identify the financial assets which constitute the main risk contributors, reallocate the portfolio, and, in this way, minimize the potential loss by minimizing the portfolio risk.

From a historical perspective, Markowitz (1952) was the first to recognize the relationship between risk and reward, and introduced the standard deviation as a proxy for risk. The standard deviation is not a good choice for a risk measure because it penalizes symmetrically both the negative and the positive deviations from the mean. It is an uncertainty measure and cannot account for the asymmetric nature of risk, i.e., risk concerns only loss. The deficiencies of the standard deviation as a risk measure were acknowledged by Markowitz who was the first to suggest the semi-standard deviation as a substitute, Markowitz (1959).

A risk measure which has been widely accepted since 1990s is the value-at-risk (VaR). In the late 1980s, it was integrated by JP Morgan on a firm-level into its risk management system. In this system, JP Morgan developed a service called RiskMetrics which was later spun off into a separate company called RiskMetrics Group. It is generally thought that JP Morgan first formulated the VaR measure. In fact, similar ideas had been used by large financial institutions. For more information about risk measures, the reader is referred to Rachev et al. (2008) and the references therein.

Although VaR has been widely adopted as a standard risk measure in the financial industry, financial professionals have identified a number of deficiencies. One important deficiency is that VaR cannot always account for the risk diversification effect. There are examples in which the portfolio VaR is larger than the sum of the VaRs of the portfolio constituents. Another important deficiency is that VaR is not informative about the extreme losses beyond the VaR level. A risk measure which lacks these deficiencies is the average value-at-risk (AVaR). It is defined as the average VaR beyond a given VaR level. Not only does it have an intuitive definition, but there are also convenient ways of computing and estimating it. As a result, AVaR is a superior alternative to VaR and is suitable for management of portfolio risk and optimal portfolio problems. The average of VaRs is computed through the integral,

$$\mathrm{AVaR}_{\varepsilon}(X) = \frac{1}{\varepsilon}\int_{0}^{\varepsilon} \mathrm{VaR}_{p}(X)\,\mathrm{d}p,$$

where ε denotes the tail probability and $\mathrm{VaR}_p(X) = -\inf\{x : P(X \leq x) \geq p\}$ is the VaR of X at tail probability p. For additional information about AVaR, see Rachev et al. (2008). If the distribution of X is absolutely continuous, then the notion of AVaR coincides with the expected tail loss (ETL) defined through the conditional expectation,

$$\mathrm{ETL}_{\varepsilon}(X) = -E\left(X|X < -\mathrm{VaR}_{\varepsilon}\left(X\right)\right).$$

For this reason, in stable Paretian models for asset returns distributions, we can use both terms interchangeably. However, even though from a mathematical viewpoint both terms are equivalent for absolutely continuous distributions, we choose ETL when combining with stable distributions for asset returns modeling since ETL is intuitively linked to the tail behavior which is a central notion in stable Paretian distributions.

We summarize the attractive properties of AVaR below:

- AVaR gives an informed view of losses beyond VaR.
- AVaR is a convex, smooth function of portfolio weights, and is therefore attractive to optimize portfolios (see Rockafellar and Uryasev 2000).
- AVaR is sub-additive and satisfies a set of intuitively appealing coherent risk measure properties (see Artzner et al. 1999).
- AVaR is a form of expected loss (i.e., a conditional expected loss) and is a very convenient form for use in scenario-based portfolio optimization. It is also quite a natural risk-adjustment to expected return (see STARR, or stable tail adjusted return ratio).

Even though AVaR is not widely adopted, we expect it to become an accepted risk measure as portfolio and risk managers become more familiar with its attractive properties. For portfolio optimization, we recommend the use of Stable distribution ETL (SETL), and limit the use of historical, normal, or stable VaR to required regulatory reporting purposes only. Finally, organizations should consider the advantages of SETL for risk assessment purposes and nonregulatory reporting purposes.

17.4 Risk Decomposition Based on SETL

The concept of SETL allows for scenario-based risk decomposition which is similar to the standard deviation-based percentage contribution to risk (PCTR). The practical issue is to identify the contribution of each position to portfolio risk. Since ETL is a tail risk measure, percentage contribution to ETL allows one to build a framework for tail risk budgeting. The approach largely depends on one of the properties of coherent risk measures given in Artzner et al. (1999), which is the positive homogeneity property

$$\mathrm{ETL}_\varepsilon(aX) = a\mathrm{ETL}_\varepsilon(X), \quad a > 0.$$

According to Euler's formula, ETL can be expressed in terms of a weighted average of the partial derivatives with respect to portfolio assets assuming that there exists a small cash account,

$$\mathrm{ETL}_\varepsilon(w'r) = \sum_i w_i \frac{\partial \mathrm{ETL}_\varepsilon(w'r)}{\partial w_i}.$$

The cash account is used to finance the infinitesimal increase of portfolio holdings in order to compute the partial derivatives of the risk measure. The left-hand side of the equation equals the total portfolio risk and if we divide both sides by it, we obtain the needed tail risk decomposition,

$$1 = \sum_i \frac{w_i}{\mathrm{ETL}_\varepsilon(w'r)} \frac{\partial \mathrm{ETL}_\varepsilon(w'r)}{\partial w_i}.$$

The same method can be applied if there is an underlying factor model in order to determine the factor percentage contribution to tail risk or, on a more general level, the systematic and nonsystematic percentage contribution. Furthermore, the partial derivatives of ETL can be computed from scenarios (see Zhang and Rachev 2006).

17.5 Portfolio Optimization with SETL

The solution of the optimal portfolio problem is a portfolio that minimizes a given risk measure, provided the expected return is constrained by some minimal value R.

In our framework, we adopt ETL as a risk measure:

$$\begin{aligned}
&\min_{w} \mathrm{ETL}_{\varepsilon}(w'r - r_{\mathrm{b}}) \\
&\text{s.t.} \\
&w'Er - Er_{\mathrm{b}} \geq R \\
&l \leq Aw \leq u,
\end{aligned} \tag{17.2}$$

where the vector notation $w'r$ stands for the returns of a portfolio with composition $w = (w_1, w_2, \ldots, w_n)$, l is a vector of lower bounds, A is a matrix, u is a vector of upper bounds, and r_{b} is some benchmark (which could be set equal to zero). The set represented by the double linear inequalities in matrix notation $l \leq Aw \leq u$ and includes all feasible portfolios.

If the benchmark is zero, $r_{\mathrm{b}} = 0$, and instead of ETL we use the standard deviation, which is an uncertainty measure. Then the optimization problem transforms into the classical Markowitz problem. Optimal portfolio problems with a benchmark are called *benchmark tracking problems*. The benchmark could be nonstochastic or stochastic, for example the return of another portfolio or a market index. In case r_{b} is nonzero and we use the standard deviation instead of ETL, the problem transforms into the classical tracking error problem.

The set of all solutions of (17.2), when varying the value of the constraint, is called the *efficient frontier*. Along the efficient frontier, there is a portfolio that provides the maximum expected return per unit of risk; that is, this portfolio is a solution to the optimal ratio problem

$$\begin{aligned}
&\max_{w} \frac{w^T Er - Er_{\mathrm{b}}}{\mathrm{ETL}_{\varepsilon}(w^{\mathrm{T}}r - r_{\mathrm{b}})} \\
&\text{s.t.} \\
&l \leq Aw \leq u.
\end{aligned} \tag{17.3}$$

An example of a reward-risk ratio is the celebrated Sharpe ratio or the information ratio depending on whether the benchmark is stochastic. In both cases, the standard deviation is used instead of the ETL. Besides the Sharpe ratio or the information ratio, many more examples can be obtained by changing the risk and, possibly, the reward function (see Biglova et al. 2004 for an empirical study).

Problem (17.3) can be transformed into a simpler problem on the condition that the risk measure is strictly positive for all feasible portfolios

$$\begin{aligned}
&\min_{x,t} \mathrm{ETL}_{\varepsilon}(x^{\mathrm{T}}r - tr_{\mathrm{b}}) \\
&\text{s.t.} \\
&x^{\mathrm{T}}Er - tEr_{\mathrm{b}} = 1 \\
&tl \leq Ax \leq tu
\end{aligned} \tag{17.4}$$

where t is an additional variable. If (x_0, t_0) is a solution to (17.4), then $w_0 = x_0/t_0$ is a solution to (17.3). There are other connections between problems (17.3) and (17.4), see Stoyanov et al. (2007) for further details.

Following the approach in Rockafellar and Uryasev (2000), (17.2) can be solved by a linear programming problem,

$$\begin{aligned}
&\min_{w,\theta,d} \ \theta + \frac{1}{N\varepsilon}\sum_{k=1}^{N} d_k \\
&\text{s.t.} \\
&w^{\mathrm{T}} Er - Er_{\mathrm{b}} \geq R \\
&-w^{\mathrm{T}} r^k + r_b^k - \theta \leq d_k,\ k = 1, N \\
&d_k \geq 0,\ k = 1, N \\
&l \leq Aw \leq u,
\end{aligned} \tag{17.5}$$

where $(r^1, \ldots, r^N)$ and $\left(r_{\mathrm{b}}^1, \ldots, r_{\mathrm{b}}^N\right)$ are scenarios for the asset returns and the benchmark generated according to the generalized stable distribution framework. See also Rachev et al. (2008) for the geometric interpretations and further information on computational complexity.

17.6 Performance Measures

The celebrated *Sharpe ratio* for a given portfolio p is defined as follows:

$$\mathrm{SR_p} = \frac{\mathrm{ER_p} - r_{\mathrm{f}}}{\sigma_{\mathrm{p}}},$$

where $\mathrm{ER_p}$ is the portfolio expected return, σ_{p} is the portfolio return standard deviation as a measure of portfolio risk, and r_{f} is the risk-free rate. While the Sharpe ratio is the single most widely used portfolio performance measure, it has several disadvantages as the standard deviation is used as a proxy for risk measure:

- σ_{p} is a symmetric measure that does not focus on downside risk.
- σ_{p} is not a coherent measure of risk (see Artzner et al. 1998).
- σ_{p} has an infinite value for non-Gaussian stable distributions.

Two alternative performance measures, consistent with the SETL framework, can be constructed, see Rachev et al. (2007). One of them, the stable tail adjusted return ratio (STARR) defined as

$$\mathrm{STARR} = \frac{w' Er - r_{\mathrm{f}}}{\mathrm{ETL}_{\varepsilon}(w' r)},$$

calculates the portfolio excess return per unit of downside risk measured by the ETL. The other performance measure is the Rachev ratio (R-ratio),

$$R\text{−ratio} = \frac{\mathrm{ETL}_{\varepsilon_1}(w'(r_\mathrm{f} - r))}{\mathrm{ETL}_{\varepsilon_2}(w'(r - r_\mathrm{f}))},$$

where ε_1 and ε_2 are two different tail probabilities and $r - r_\mathrm{f}$ is the vector of asset excess returns. The R-ratio is a generalization of the STARR. Choosing appropriate levels of ε_1 and ε_2 in optimizing the R-ratio, the investor can seek the best risk/return profile for her portfolio. For example, an investor with portfolio allocation maximizing the R-ratio with $\varepsilon_1 = \varepsilon_2 = 0.01$ seeks an exceptionally high return and protection against high losses.

17.7 An Empirical Example

Racheva-Iotova and Stoyanov (2006) provide a back-testing example of a long-only optimal portfolio strategy using the Russell 2000 universe. The back-testing time period is 10 years (December 1993–December 2004) with a monthly frequency. In the optimization algorithm, the proprietary stable model in Cognity Risk and Portfolio Optimization System is employed in which the SETL methodology is implemented. In the strategies, the Russell 2000 index is used as the benchmark; that is r_b is the return on the Russell 2000.

The optimization constraints are the following:

- 0–3% limit on a single stock
- ±3% industry exposure with respect to the benchmark; the industries being defined by Ford Equity Research
- The active return is strictly positive
- The two-way turnover is below 15% per month. This constraint is used as a soft constraint (i.e., may exceed at times). Also, no limit is imposed in July because the benchmark is adjusted in July.

The back-testing is performed in the following way. They use 450 stocks as the initial universe. One year of daily data is used to calibrate the model and monthly scenarios are produced by it. Then a version of the optimal portfolio problem (17.8) is solved in which a tail probability of 5% is selected for the ETL. At the end of the month, the present value of the portfolio is calculated. The process is repeated for the next month. Figure 17.2 shows the present values of the stable ETL portfolio compared to the Russell 2000 index.

In addition to the stable method, monthly back-testing is performed for a version of the Markowitz problem (17.6). Racheva-Iotova and Stoyanov (2006) use a factor model comprising eight factors and five years of monthly data to calibrate it. Each month, the covariance matrix is estimated through the factor model and the optimization problem is solved. The present value of the portfolio is calculated at the end of the month. Figure 17.3 shows the evolution of the portfolio's present values.

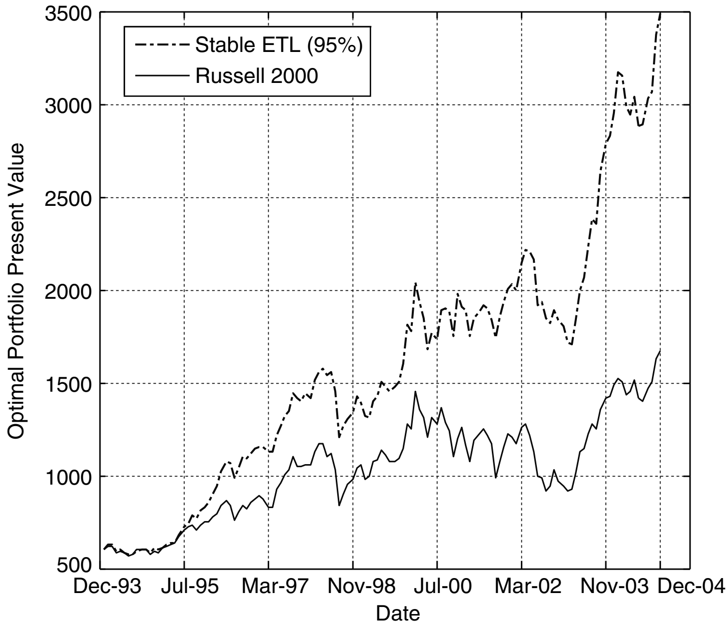


Fig. 17.2 The time evolution of the present values of the stable ETL portfolio compared to the Russell 2000 index (reproduced from Fig. 1 in Racheva-Iotova and Stoyanov 2006)

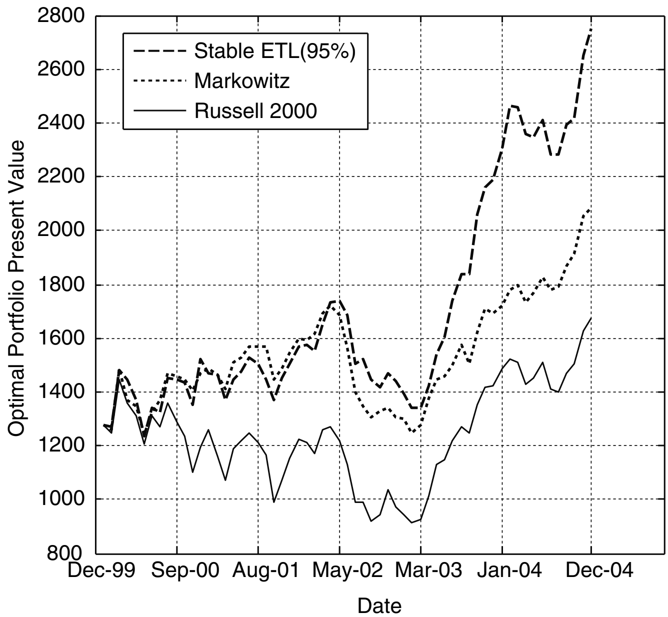


Fig. 17.3 The time evolution of the present values of the Markowitz and the stable ETL (*scaled*) portfolios compared to the Russell 2000 index (reproduced from Fig. 2 in Racheva-Iotova and Stoyanov 2006)

Table 17.1 Average number of holdings (reproduced from Table 1 in Racheva-Iotova and Stoyanov 2006)

	Stable ETL	Markowitz
10 year	112	
5 year	105	137
3 year	102	110
2 year	100	104
1 year	104	100

Table 17.2 Average monthly turnover (reproduced from Table 2 in Racheva-Iotova and Stoyanov 2006)

	Stable ETL (%)	Markowitz (%)
11 months	16	18
July	163	85
All months	27	24

Table 17.3 Annualized information ratio (reproduced from Table 3 in Racheva-Iotova and Stoyanov 2006)

	Stable ETL	Markowitz
10 year	0.74	
5 year	0.71	0.29
3 year	0.93	−0.24
2 year	0.74	−0.57
1 year	1.22	1.03

Table 17.4 Sharpe ratios (reproduced from Table 3 in Racheva-Iotova and Stoyanov 2006)

	Stable ETL	Markowitz	Russell 2000
10 year	1.01		0.42
5 year	0.92	0.68	0.36
3 year	1.22	0.71	0.58
2 year	2.13	1.99	1.82
1 year	1.66	2.16	1.19

Note that the present value of the stable portfolio is scaled so as to start with the same capital as the Markowitz model.

Additional information is given in Tables 17.1 and 17.2. The average monthly turnover is defined as the dollar-weighted purchases plus the dollar-weighted sales. Tables 17.3 and 17.4 provide details on return-risk ratios. The information ratio is the active return per unit of tracking error.

17.8 Conclusion

In this chapter, we described and discussed the SETL framework implemented in the Cognity Risk Management and Portfolio Optimization product. The SETL framework is appealing because it is based on realistic assumptions about asset return distributions, incorporates a downside risk measure, and can be used for risk budgeting and portfolio optimization. With the help of empirical examples, we demonstrated that the SETL framework is more realistic than the traditional models based on the normal distribution and may lead to better performance.

References

Akgiray, A. and G. Booth (1988) The stable-law model of stock returns, *Journal of Business and Economic Statistics*, 6, 51–57.

Akgiray, A. and G. Lamoureux (1989) Estimation of stable-law parameters: A comparative study, *Journal of Business and Economic Statistics*, 7, 85–93.

Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath (1999) Coherent measures of risk, *Mathematical Finance*, 6, 203–228.

Bibby, B.M. and M. Sorensen (2003) Hyperbolic processes in finance, in *Handbook of heavy-tailed distributions in finance*, S. Rachev (ed.), pp. 212–248.

Biglova, A., S. Ortobelli, S. Rachev, and S. Stoyanov (2004) Different approaches to risk estimation in portfolio theory, *Journal of Portfolio Management*, 31, 103–112.

Bradley, B.O. and M.S. Taqqu (2003) Financial risk and heavy tails, in *Handbook of heavy tailed distributions in finance*, S.T. Rachev (ed.), Elsevier/North-Holland, Amsterdam.

Embrechts, P., F. Lindskog, and A. McNeil (2003) Modelling dependence with copulas and applications to risk management, in *Handbook of heavy-tailed distributions in finance*, S.T. Rachev (ed.), pp. 329–384.

Fama, E. (1963) Mandelbrot and the stable Paretian hypotheis, *Journal of Business*, 36, 420–429.

Fama, E. (1965) The behavior of stock market prices, *Journal of Business*, 38, 34–105.

Janicki, A. and A. Weron (1994) *Simulation and chaotic behavior of alpha-stable stochastic processes*, Marcel Dekker, New York.

Mandelbrot, B. (1963) The variation of certain speculative prices, *Journal of Business*, 7(1), 77–91.

Markowitz, H.M. (1952) Portfolio selection, *Journal of Finance*, 7(1), 77–91.

Markowitz, H.M. (1959) *Portfolio selection: Efficient diversification of investments*, Wiley, New York.

Racheva-Iotova, B. and S. Stoyanov (2006) Post-modern approaches for portfolio optimization, forthcoming in the *Handbook on IT and finance*, Springer, Berlin.

Rachev, S. (2003) *Handbook of heavy-tailed distributions in finance*, Rachev (ed.), Elsevier/North-Holland, Amsterdam.

Rachev, S. and S. Mittnik (2000) *Stable Paretian models in finance*, Wiley, New York, Series in Financial Economics.

Rachev, S., F. Fabozzi, and C. Menn (2005) *Fat tailed and skewed asset return distributions*, Wiley, New York, Finance.

Rachev, S., S. Stoyanov, A. Biglova, and F. Fabozzi (2006) An empirical examination of daily stock return distributions for U.S. stocks, in *Data analysis and decision making*, Springer, Berlin.

Rachev, S., D. Martin, B. Racheva, and S. Stoyanov (2007) Stable ETL optimal portfolios and extreme risk management, in *Risk assessment: decisions in banking and finance*, G. Bol, S. Rachev, and R. Wureth (eds.), Springer, Berlin.

Rachev, S., S. Stoyanov, and F. Fabozzi (2008) *Advanced stochastic models, risk assessment, and portfolio optimization: the ideal risk, uncertainty, and performance measures*, Wiley, Finance.

Rockafellar, R.T. and S. Uryasev (2002) Conditional value-at-risk for general loss distributions, *Journal of Banking and Finance*, 26(7), 1443–1471.

Samorodnitsky, G. and M. Taqqu (1994) *Stable non-Gaussian random processes*, Chapman & Hall, New York, London.

Sklar, A. (1996) Random variables, distribution functions, and copulas – a personal look backward and forward, in Ruschendorff et al. (eds.) *Distributions with fixed marginals and related topics*, Institute of Mathematical Sciences, Hayward, CA.

Stoyanov, S., S. Rachev, and F. Fabozzi (2007) Optimal financial portfolios. *Applied Mathematical Finance*, 14(5), 401–436.

Zhang, Y. and S. Rachev (2006) Risk attributions and portfolio performance measurements, *Journal of Applied Functional Analysis*, 4(1), 373–402.

Part III
Applications of Portfolio Construction, Performance Measurement, and Markowitz Data Mining Corrections Tests

Chapter 18
Linking Momentum Strategies with Single-Period Portfolio Models

John M. Mulvey, Woo Chang Kim, and Mehmet Bilgili

18.1 Introduction

Since the early work of Markowitz (1952, 1956), the single-period mean-variance model has been the norm in portfolio management. Over the years, numerous variants of the core model have been proposed and implemented to improve performance in practical settings. However, empirical tests suggest that investment performance has not been outstanding in many cases. In this paper, we illustrate examples to increase the investment performance of such models by actively exploiting a significant market anomaly – momentum effects.

The most persistent equity anomaly involves the predictability of stock returns based on the past performance, which is often referred to as the momentum effect. The typical investment strategy in academic articles that exploits the effect is to buy winners and sell losers based on the intermediate term performance (3 to 12 months) proposed by Jegadeesh and Titman (1993). Many papers such as Cleary and Inglis (1998), Rouwenhorst (1998), Kang et al. (2002), and Demir et al. (2004) have documented that such a strategy is profitable, except a few stock markets Liu and Lee (2001) and Hameed and Kusnadi (2002). Further, money managers, dominant players in the stock market, are reported not only to employ the momentum effect, but also to improve their performance by applying it. For instance, Grinblatt et al. (1995), Nofsinger and Sias (1999), Sias et al. (2001), Badrinath and Wahal (2002), Sapp and Tiwari (2004), and Sias (2004) document that a significant proportion of active funds adopt the momentum strategy as their equity selection rules. Carhart (1997) illustrates that the performance persistence of mutual funds can be explained by the 1-year momentum effect. Recently, Mulvey and Kim (2008a, b) show that the active equity funds in the US share similar performance patterns with the industry-level long-only momentum strategy, and the similarity is stronger for the funds with good performance. Thus, we apply the specialized momentum strategy as a basis, and benchmark for single-period optimization models in the equity domain.

J.M. Mulvey (✉), W.C. Kim, and M. Bilgili
Princeton University, Princeton, NJ, USA

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_18,
© Springer Science+Business Media, LLC 2010

Unlike traditional approaches, we adopt industry-level data for the empirical analysis. Why do we employ industry-level data? First, compared to stock-level analysis, it reduces idiosyncratic risks. Since the mean-variance models require the estimated market parameters, it may lead to unstable test results to adopt stock-level data without employing specialized parameter estimation techniques. In contrast, one can readily eliminate such issues by analyzing the broad asset classes like industries. Second, the strategy is becoming easy to implement due to the introduction of various exchange traded funds (ETFs). In addition, the industry-level momentum strategy has displayed outstanding performance (Mulvey and Kim 2007). Importantly, one can obtain better diversification benefits from industry-wide market segmentation, as compared to the current size/style break-outs. We shall discuss the details of the issue in the later section.

The main objectives of this paper are as follows: (1) The Markowitz model requires estimating parameters for return distributions of the assets, and these are often hard to estimate. We employ momentum patterns to see if they help with a Markowitz model and several variants. (2) In addition, we compare the performances of several popular mean-variance models in various settings.

The remainder of the paper is organized as follows. In section II, we briefly discuss four asset allocation models employed in this paper – traditional Markowitz, Black-Litterman, Grauber-Hakansson, and robust optimization models. In the following section, we illustrate how the industry-wide equity market segmentation can provide better diversification benefits. Empirical results and conclusions follow.

18.2 Models: Markowitz Model and its Variants

In spite of its popularity, several issues arise regarding the practical implementation of the Markowitz model. A major issue is its sensitivity to input changes. Since the optimal portfolio from the mean-variance approach is chosen among the extreme points of the feasible region, small changes in the estimated parameters of the market distribution can lead to radically different optimal points. As a consequence, relatively small errors in the parameter estimation can potentially cause a steep decrease of investment performance. Such a high sensitivity is undesirable for practical applications.

Many models have been proposed to overcome this shortcoming. A popular approach is to utilize robust estimators for the mean and the variance. Instead of using the unbiased estimators for the market distribution, one can reduce the estimation error by shrinking the sample mean and the sample covariance toward structured estimators. Such shrinkage methods are well documented in Jobson and Korkie (1981), Jorion (1986), Pasotr (2000), and Larsen and Resnick (2001). Also, in a similar context, Black and Litterman (1990) propose a model to blend the investor's view with the market estimators. In their model, the investor's view, which is represented as a linear relation among the expected returns of the individual assets, is mixed to the market equilibrium via a Bayesian approach. See Satchell and Scowcroft (2000), and Idzorek (2004) for the detailed discussion.

An alternative approach is the portfolio re-sampling technique. In this approach, market parameters are re-sampled via Monte Carlo simulation, and the portfolio weights are obtained by averaging the optimal solutions of individual mean-variance problems with the generated estimators. The Michaud model (1998), for example, generates random samples from the estimated mean and variance, and obtains a new set of market parameter estimators from samples. The efficient frontier corresponding to this simulation is produced by minimizing a set of evenly spaced portfolio variances. After repeating the procedure sufficiently often, one can gain the re-sampled portfolio weight by averaging the optimal allocations with the same variance ranking. In a sense, this approach addresses the sensitivity issue by averaging the outputs from perturbed samples, while the robust estimator methods smoothen the estimators.

There also has been a significant effort to improve the robustness of the optimal portfolio allocation by analytically reflecting the uncertainty of estimated parameters with the help of convex analysis. These models typically define the uncertainty set for the parameters and formulate the optimal allocation as a convex optimization problem to consider the worst case. Accordingly, this approach is referred to as a robust optimization model. For instance, when the mean return or the covariance is relaxed to lie in an ellipsoid, the mean-variance problem can be rewritten as a second order cone programming (SOCP) problem. If the covariance takes a finite number of matrices, it is formulated as a quadratic constrained quadratic programming (QCQP) problem. Note that both SOCP and QCQP are convex programs which can be solved efficiently. See Ben-Tal and Nemirovski (1995), El Ghaoui and Lebret (1997), Ben-Tal and Nemirovski (2001), and Boyd and Vandenberghe (2004) for further details.

We pick one model for each of the approaches as well as the traditional Markowitz model to evaluate their historical performance. Black-Litterman model, Grauer-Hakansson model (1985), and SOCP model for the mean relaxation are chosen for the robust parameter estimation, re-sampled portfolio, and robust optimization approaches, respectively. Meucci (2005) and Fabozzi et al. (2007) discuss these models as well as other approaches in detail.

18.2.1 Robust Estimator: Black-Litterman

We employ a simplified version of Black-Litterman model described in Meucci (2007). With the normality assumption, let the prior on the market be μ and Σ. That is, the n-dimensional random return vector

$$r \sim N(\mu, \Sigma)$$

Also, let the view on the market be expressed as the following linear function.

$$v = P\mu + \in,$$

where P is k-by-n matrix corresponding to k views on the market along with k-vector v, and $\in$ is the error term that follows $N(0, \Omega)$. Note that Ω represents the investor's confidence on the view. For simplicity, we set

$$\Omega = \left(\frac{1}{c} - 1\right) P\Sigma P^T.$$

For this specific choice of the uncertainty matrix Ω, c determines the confidence: values in Ω decrease as c increases from 0 to 1, which causes decrease in the variance for the view, thus making it more certain. For our tests, we use 0.01, 0.5 and 0.99 for the values of c. Then, the Black-Litterman estimators for the expected return and the covariance can be shown as follows:

$$\mu_{\mathrm{BL}} = \mu + \Sigma P^T (P\Sigma P^T + \Omega)^{-1}(v - P\mu),$$

and

$$\Sigma_{\mathrm{BL}} = \Sigma - \Sigma P^T (P\Sigma P^T + \Omega)^{-1} P\Sigma.$$

For the prior, μ and Σ, we employ the sample mean and the sample covariance of long-term (5 years) historical returns at each time period. P and v, which represent the investor's view on the market, are constructed to reflect the performance persistence of the recent winners, or the momentum effect. They are chosen in such a way that the average value of the expected returns of recent top 10% winners is higher than the remaining 90% by an arbitrary amount v in annualized return. In other words, for the index set I of the recent top 10% winners, and n_w and n_l, the number of winners and losers, respectively,

$$\frac{1}{n_w}\sum_{i\in I}\mu_i = \frac{1}{n_l}\sum_{j\in I}\mu_j + v.$$

Therefore, P is 1-by-n vector, where

$$P_i = \frac{1}{n_w}\text{for } i \in I, \quad \text{and} \quad P_j = \frac{1}{n_l}\text{for } j \notin I.$$

We choose 3-, 6-, 9-, 12-, 24-, 36- and 60-month evaluation time lengths to obtain P and 1% for v. Once all input parameters are set, the mean-variance approach is employed to obtain the optimal portfolio allocation. Note that this approach is consistent with the long-only momentum strategies proposed in Mulvey and Kim (2007). They construct the long-only industry-level momentum portfolios by holding recent top 10% winner industries with equal weights in several stock markets. The portfolios have outperformed the benchmark market indices in most of the tested markets.

18.2.2 Re-sampled Portfolio: Grauer-Hakansson

Several authors have successfully implemented a sequence of single-period optimization model based on optimizing a Von Neumann-Morgenstern (VM) expected utility function. See, for example, Grauer and Hakansson (1985), and Mulvey et al. (2006). Also, Markowitz (1952) discusses the advantages of employing the model. Herein, we implement the model via an iso-elastic utility function:

$$u(w) = \frac{1}{\gamma} w^{\gamma}, \quad \text{where } \gamma \leq 1.$$

VM model is based on the one-step tree representation of scenarios as illustrated in Fig. 18.1. Given the wealth W_t at time t, the future wealth at time $t + 1$ for scenario s_i is $W_t(1 + R_{s_i})$ with a probability of π_{s_i}, where R_{s_i} is the portfolio return between t and $t + 1$. That is,

$$R_{s_i} = \Sigma_j W_j r_{j,s_i},$$

where w_j is the weight on asset j and r_{j,s_i} is the return of asset j for scenario s_i. We use each monthly observation as one scenario and set the probability equal across scenarios. For instance, when T-month historical data is employed for the scenario construction, the asset returns at month i are set to be the scenario return for scenario s_i, and $1/T$ is assigned as the probability for each scenario, thus having T scenarios in total.

The optimal portfolio weight is calculated from the following expected utility maximization problem with a nonnegativity constraint on the portfolio weights, since this model does not utilize the mean-variance approach.

$$\text{Max}_{w \geq 0} \mathbb{E} U(W_{t+1}) = \frac{1}{\gamma} \sum_{i=1}^{T} \frac{W_{t+1,s_i}^{\gamma}}{T} = \frac{1}{\gamma T} \sum_{i=1}^{T} W_{t+1,s_i}^{\gamma} \quad \text{for } \gamma \leq 1$$

An efficient frontier is generated by varying the risk-aversion coefficient (γ); $\gamma = 1$ which corresponds to the risk neutral case, and the resulting portfolio becomes more conservative as γ decreases. As in the Black-Litterman model, we adopt 3-, 6-, 9-, 12-, 24-, 36- and 60-month as the parameter estimation periods.

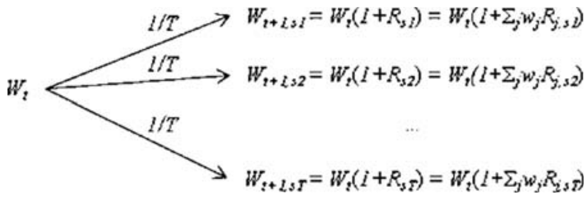


Fig. 18.1 Tree representation of the Grauer-Hakansson model

18.2.3 Robust Optimization: SOCP for Ellipsoidal Relaxation of μ

Consider the following mean-variance problem.

$$\begin{array}{ll} \text{Maximize} & w^T \mu \\ \text{Subject to} & w^T \Sigma w \le \sigma_{\text{target}}^2 \\ & 1^T w = 1, \quad w \ge 0 \end{array}$$

Suppose the parameter μ is uncertain, but is known to lie in an ellipsoid induced by $\bar{\mu}$ and P:

$$\mu \in \varepsilon := \{\bar{\mu} + Pu | d_2(u) \le q\}, \text{ where } d_2 \text{ is Euclidean norm}$$

Note that the size of the ellipsoid increases as q gets large, and so q can be interpreted as the degree of the uncertainty on μ.

Now suppose we strive to obtain the robust asset allocation in a sense that the solution is optimal under the worst-case scenario. Then the mean-variable problem can be restated as follows:

$$\begin{array}{ll} \text{Maximize} & \inf\limits_{\mu \in \varepsilon} w^T \mu \\ \text{Subject to} & w^T \Sigma w \le \sigma_{\text{target}}^2 \\ & 1^T w = 1, \quad w \ge 0 \end{array}$$

Since for every given w,

$$\begin{aligned} \inf_{\mu \in \varepsilon} w^T \mu &= w^T \bar{\mu} + \{Pu | d_2(u) \le q\} = w^T \bar{\mu} - q d_2(P^T w) \\ &= w^T \bar{\mu} - q \sqrt{w^T P P^T w} \end{aligned}$$

we have

$$\begin{array}{ll} \text{Maximize} & w^T \bar{\mu} - q \sqrt{w^T P P^T w} \\ \text{Subject to} & w^T \Sigma w \le \sigma_{\text{target}}^2 \\ & 1^T w = 1, \quad w \ge 0 \end{array}$$

Therefore, the original problem with the ellipsoid relaxation can be expressed as

$$\begin{array}{ll} \text{Maximize} & w^T \bar{\mu} - z \\ \text{Subject to} & d_2 \left(\Sigma^{\frac{1}{2}} w \right) \le \sigma_{\text{target}} \\ & 1^T w = 1, \quad w \ge 0 \\ & q d_z(P^T w) \le z, \end{array}$$

where $\Sigma^{\frac{1}{2}}$ is a Cholesky decomposition of Σ. The final form is SOCP, which can be solved relatively easily via convex optimization techniques. Note that the additional constraint ($qd_z(P^T w) \leq z$) plays a role of keeping w from moving toward the direction to which the uncertainty increases. See Boyd and Vandenberghe (2004) and Meucci (2007).

There are three input parameters that should be estimated for the implementation: $\bar{\mu}$, Σ and P. We employ the sample mean and the sample covariance for $\bar{\mu}, \Sigma$. Also, for simplicity, we choose P in such a way that $PP^T = \text{diag}(\Sigma)$. In addition, we vary q from 0 to 1 to investigate the effect of the uncertainty level. As in the previous cases, parameters are estimated from 3-, 6-, 9-, 12-, 24-, 36- and 60-month historical data, in order to evaluate the impact of various momentum-based rules.

18.2.4 Markowitz Model and General Experiment Settings

We adopt the following traditional mean-variance problem with the nonnegativity constraint as the benchmark model.

$$\begin{aligned}
&\text{Maximize } w^T \mu \\
&\text{Subject to } w^T \Sigma w \leq \sigma^2_{\text{target}} \\
&\qquad\qquad 1^T w = 1, \quad w \geq 0
\end{aligned}$$

The sample mean and the sample covariance from 3-, 6-, 9-, 12-, 24-, 36- and 60-month historical data are employed for μ and Σ.

For all four models, we conduct the following sequential portfolio allocation: at the beginning of the evaluation period, input parameters as described earlier, and the optimal allocation are determined. Then, the assets are held for 3-, 6- or 12-month based on the optimal weights and rebalanced every month to the initial weights (i.e., fixed mix portfolios). After the holding period has ended, another set of the allocation is conducted. The procedures are repeated until the end of the sample period. Table 18.1 summarizes four models and accompanied parameters.

18.3 Data: Benefits of Industry-wide Market Segmentation

In this section, we discuss the importance of employing generic asset categories within an optimal portfolio model. Clearly, the performance of optimal asset allocation models is highly dependent upon the characteristics of the given assets. Efforts to find asset classes with good properties should precede the model selection. Current practical approaches typically divide the equity market based on the sizes and the style prospects of stocks (e.g., large-core, mid-growth, small-value, etc.). Since the size/style segmentation is simply a scheme to cut the group of investable vehicles, it is natural to ask if the current segmentation scheme can be improved.

Table 18.1 Summary for asset allocation models

Scheme	Model	Market estimators and parameters
Robust estimator	Black-Litterman	μ: sample mean from 60-month data Σ: sample covariance from 60-month data v and P: winners from 3-, 6-, 9-, 12-, 24-, 36- and 60-month data $\Omega = \left(\frac{1}{c} - 1\right) P\Sigma P^T$: confidence on investors view. ($c = 0.01$, 0.5, and 0.99)
Re-sampled portfolio	Grauer-Hakansson	$r_j s_i$: monthly return on asset j from 3-, 6-, 9-, 12-, 24-, 36- and 60-month data
Robust optimization	SOCP for ellipsoidal μ	$\bar{\mu}$: sample mean from 3-, 6-, 9-, 12-, 24-, 36- and 60-month data Σ: sample covariance from 3-, 6-, 9-, 12-, 24-, 36- and 60-month data P: diagonal elements of Σ
Markowitz model	Mean-variance model	μ: sample mean from 3-, 6-, 9-, 12-, 24-, 36- and 60-month data Σ: sample covariance from 3-, 6-, 9-, 12-, 24-, 36- and 60-month data

In this regard, Mulvey and Kim (2008a, b) illustrate that a segmentation scheme based on industry-level definitions can potentially improve the performance of investment vehicles, as compared to the current size/style scheme. First, an industry-wide segmentation can provide more consistent membership over time. The reason is clear: firms cannot easily change industries which they belong to, while their sizes and growth perspectives vary. This property not only enables investors to track each of the segmentations easily, but also potentially improve performance active funds; many mutual fund managers are typically bound to form their portfolios with stocks corresponding to the styles of the funds. Thus, fund managers may be forced to perform undesired portfolio adjustments to reflect the changes in the benchmark components, when the membership for each breakout changes. Such a procedure typically limits the fund managers' choices, which may lead to inferior investment performance.

Figure 18.2 shows how unstable the style/size classifications have been for the last decade as compared to the industry segmentation scheme. Here, we conduct style analysis over a recent period of time in a sequential fashion. The technology industry (left in Fig. 18.2) stays as growth-oriented, while its size has shrunken from large-cap to small-cap, and then grown back to large-cap. Similarly, the healthcare industry (center in Fig. 18.2) has been classified as large-cap, while its growth perspective has changed over the sample period. Also, the oil and gas industry (right in Fig. 18.2) has moved over the three quadrants of the style/size map. Since the membership of firms within an industry is stable, it is clear that the constituents of style/size market break-outs have changed frequently.

A critical benefit from the industry segmentation is improved diversification potential. Figure 18.3 depicts the average correlations of market breakouts from

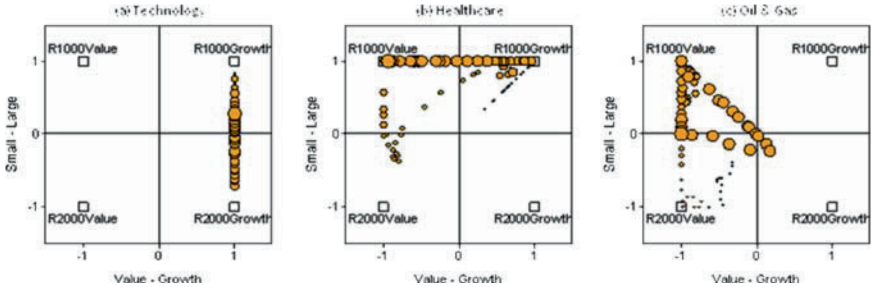


Fig. 18.2 Drift of industries in style/size map from December 1996 to November 2006. This figure illustrates drift of (**a**) technology, (**b**) healthcare, and (**c**) oil and gas industries in style/size map. The sample period is from December 1996 to November 2006. Relative position of each circle represents the size and the style of industries for each 24-month long period compared to 4 Russell indices on a rolling time basis. The circle sizes increase as time passes. The positions of the circles are based on regressions of industry returns against the Russell indices

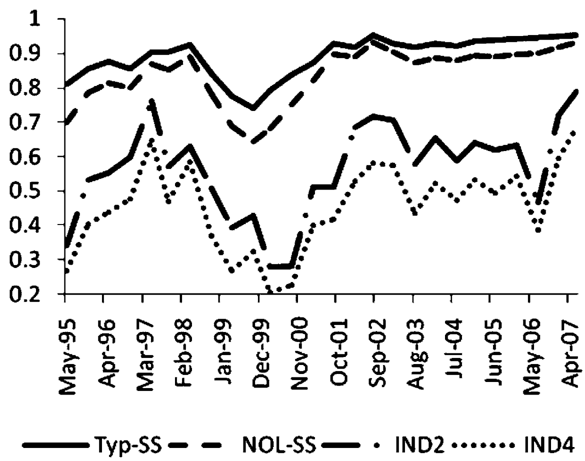


Fig. 18.3 Average correlations within different market segmentation schemes. This figure illustrates the average correlations for 4 different market breakouts defined in the top table. The sample period ranges from June 1995 to December 2007. For each of market segmentations, correlations for all possible index pairs are calculated from daily returns, and then averaged across those pairs. The unit time length is 6 months (126 trading days). See appendix for the list of Datastream sectors

different segmentation schemes. Typ-SS and NON-SS represent the size/style segmentation scheme, while IND2 and IND4 correspond to the industry break-out. The message is clear: from the same stock market, one can obtain less correlated vehicles by applying the industry-wide segmentation scheme than the conventional size/style approach. It has a critical implication to the asset allocation models; it can provide better diversification for the broad asset allocation problems.

Description	Code	Indices included
Typical style/size breakouts	Typ-SS	R1000, R1000G, R1000V, RMid, RMidG, RMidV, R2000, R2000G, R2000V
Non-overlapping style/size breakouts	NOL-SS	R200G, R200V, RMidG, RMidV, R2000G, R2000V
DataStream level 2 sectors	IND2	10 Industries indices
DataStream level 4 sectors	IND4	38 Industries indices

18.4 Test Results

In this section, the investment performance of the models introduced in section II is compared with several benchmark portfolios. To construct the portfolio corresponding to each model, we update estimators for the expected return and the covariance periodically, and employ optimization on a moving basis. There are two critical parameters: "look-back period" refers to the length of the historical data to estimate return and covariance matrices, and "holding period" shows the investment period for each asset allocation decision. For instance, at a given period of time a strategy with 3-month look-back period and 6-month holding period means that the inputs are estimated from the daily returns of past 3 months, and once the optimal allocation is obtained, the portfolio is held for 6 months. Note that all assets in a portfolio are rebalanced to their corresponding weights at the end of each month. Also, in order to eliminate the timing bias, the average returns from portfolio with different starting points are employed. The extra parameters for robust optimization and Black-Litterman have been chosen as mentioned in section II. In this context, we apply the Markowitz model and its variants to the industries as defined in Datastream. There are 38 industries with one market index, and we employ daily data from January 1976 to December 2007. Since the models require an initial period of time to estimate parameters, all constructed portfolios begin on January 1980. See the appendix for the list of industries.

18.4.1 Model Comparisons: Which Model Is Better?

The historical performance is obtained after sequentially solving the relevant optimization problems, and calculating the return series for each model (Fig. 18.4). Different risk tolerance levels are set to generate points along the line. For (a)–(f) in Fig. 18.4, the holding period is fixed to 6 months, while the look-back period are set to 3-, 6-, 12-, 24-, 36- and 60-month, respectively.

When the look-back period is 3-month, the Black-Litterman model dominates the other models (Fig. 18.4a), particularly at the points corresponding to the highly risk

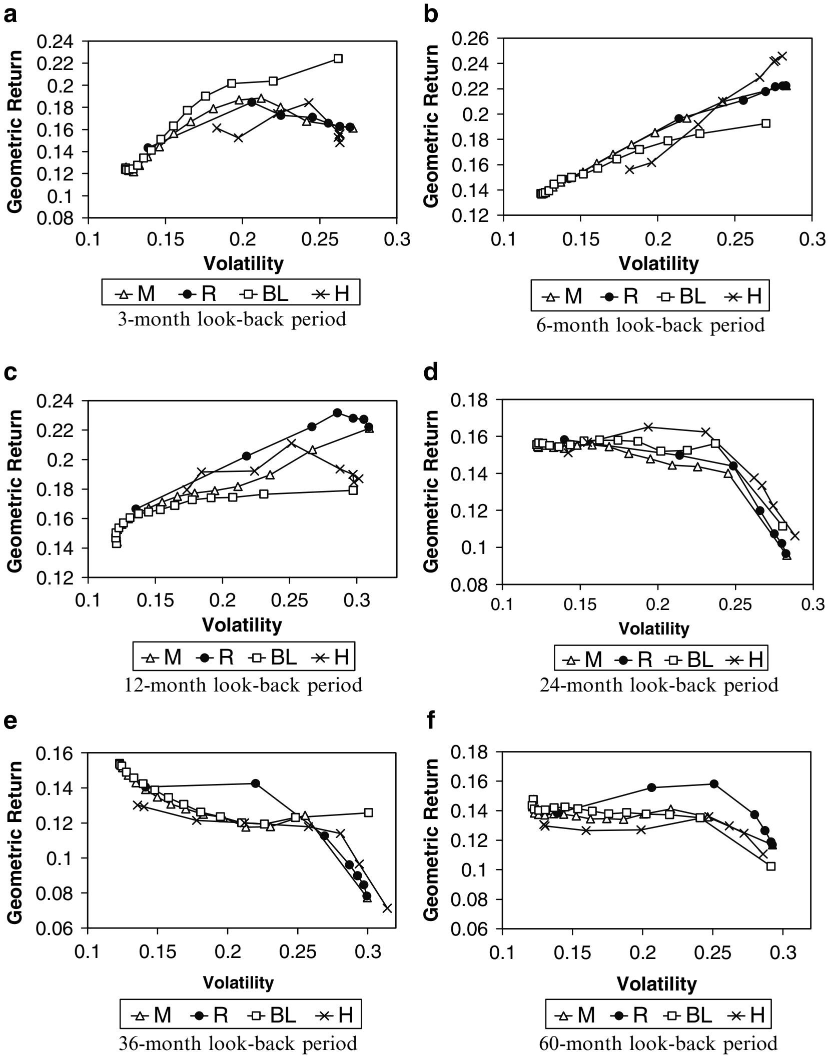


Fig. 18.4 Historical performance of 4 different models from January 1980 to December 2007. This figure illustrates the historical investment performance of the 4 different models – Markowitz (M), Black Litterman (BL), Grauber-Hakansson (H), and robust optimization (R) Models. The sample period of time is from 1980 to 2007. Holding period is set to 6-month across all three figures, while look-back periods are set to (**a**) 3-month, (**b**) 6-month, (**c**) 12-month, (**d**) 24-month, (**e**) 36-month, and (**f**) 60-month

portfolios. However, robust optimization and Grauber-Hakansson models produce better results when 12-months look-back period is adopted (From Fig. 18.4c). Furthermore, there is no dominating model in Fig. 18.4e; returns are not significantly different from each other, while none of the approaches offer a promising decrease in volatility with respect to the others. In summary, for these tests, investment performance across different models is comparable with each other, and it highly depends on the parameter settings.

There are two other meaningful observations. First, the Markowitz model performs fairly well in practical settings over the entire sample period. Second, while the ex post performance lines in the mean-variance plane preserve upward slopes and concave shapes when the look-back period is equal or less than a year, they become downward sloping when it is longer than a year. Thus, performance deteriorates as the level of the risk tolerance increases for look-back period greater than 1 year. In fact, these findings coincide with the superior performance of the momentum strategy, and the discussion follows in the next subsection. Note that the results lead to the same conclusion when different holding periods are employed (3-, and 12-month), while the investment performance generally becomes worse as the holding period gets longer.

18.4.2 Comparisons along Different Look-back Periods: Blending in Momentum Effects

In this subsection, we evaluate several look-back periods. Figure 18.5 illustrates performance for different look-back periods (3-, 6-, 9-, 12-, 24-, 36-, 60-month) with a 6-month holding period. For all models, portfolios with short term look-back periods (3-, 6-, 9-, 12-month) show superior performance to ones with longer look-back periods (24-, 36-, and 60-month). Furthermore, the lines change shape from being upward sloping and concave to downward sloping and convex as the look-back period gets longer. As pointed out in the previous subsection, it implies that taking higher risk actually reduces the ex post returns. Tests with different holding periods (3-, and 12-month) yield similar results.

These results imply that estimating market parameters from short look-back periods are better than ones from longer look-back periods. All models rely on the assumption that estimated returns and covariance matrices are proxies for the future values of these parameters. So, their performance depends on the persistence of these estimates along time. Since industries with better performance during the look-back periods would have higher weights in the portfolio for the subsequent period regardless of the model choice, it is evident that recent data provides better forecast on the distribution of the future returns.

Why is the shorter holding period better? The answers can be readily found from the equity price momentum effects: empirical studies suggest that winner stocks for the past 3 to 12 months outperform loser stocks for the following 3 to 12 months and show worse performance after 3 to 5 years. Therefore, when shorter holding periods are employed, the models put more weights on the recent winners, leading to a successful blend in the momentum effects with optimal asset allocation models. In contrast, the portfolios from longer look-back period bet against the momentum effects, which would potentially cause inferior investment performance.

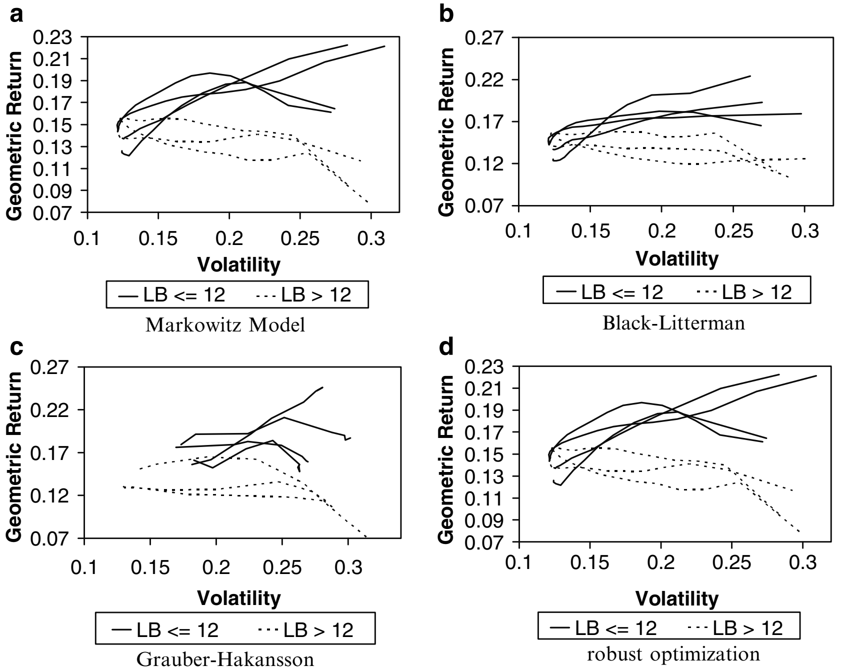


Fig. 18.5 Performance across different look-back periods. This figure illustrates the historical investment performance of each of the 4 different models with different look-back periods (LB). The sample period of time is from 1980 to 2007. Holding period is set to 6 months across all three figures, while look-back periods are set to 3-, 6-, 9-, 12-, 24-, 36- and 60-month. (**a**) Markowitz Model (**b**) Black-Litterman (**c**) Grauber-Hakansson and (**d**) robust optimization

One significant question remains: Is the momentum effect strong enough to improve investment performance of the asset allocation models? To see this, we compare the historical performance of the models to three benchmark portfolios – the market index, 60–40 fixed mix portfolio, and the long-only momentum portfolio (Fig. 18.6). The 60–40 mix is constructed by investing 60% of the wealth to market index, and 40% to treasury bills with monthly rebalancing. Also, the performance of the long-only momentum portfolio is obtained by holding the winner industries based on the past 3-, 6-, 9-, and 12-month returns for 6 subsequent months. The chosen industries are equally weighted and rebalanced every month. For the optimal asset allocation portfolios, 6-month look-back period and 6-month holding period are employed; this setting provides relatively good investment performance for all four models.

Figure 18.6 provides a clear answer for the question. The performance of long-only momentum strategy lies on the best performance line of the optimal asset allocation models in the mean-variance plane for the entire sample period, meaning that the simple momentum rule has performed equivalently, if not better, to the optimization models. This implies that the momentum effects have been significant, and they can improve investment performance, if investors utilize the effects properly. It is interesting to note that the momentum strategy shows stronger performance

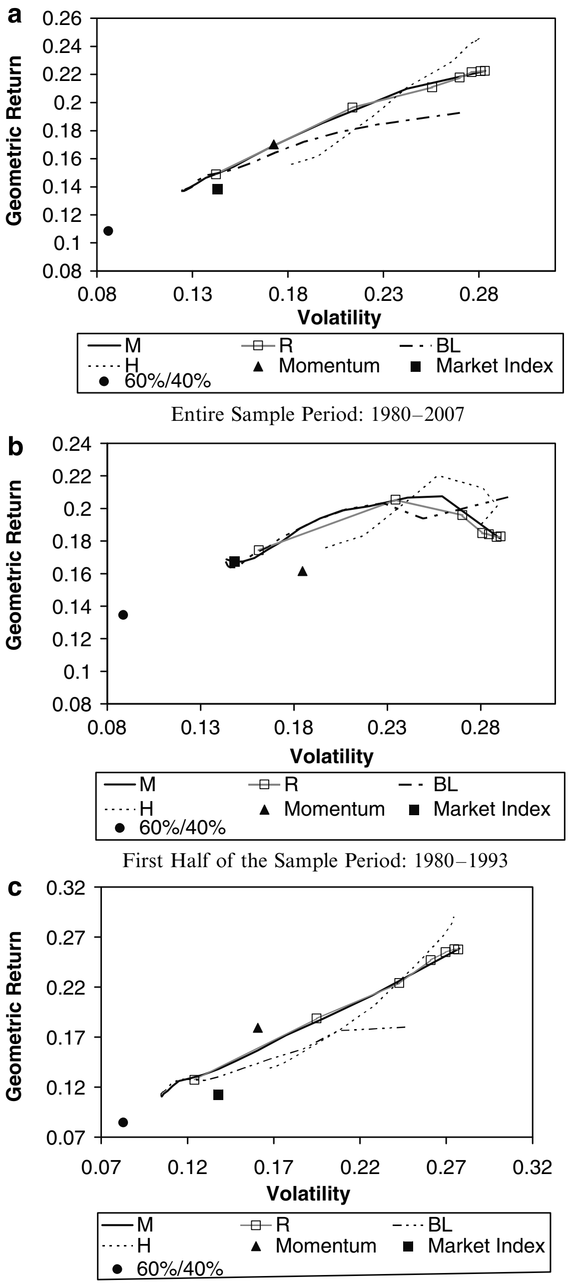


Fig. 18.6 Performance comparisons to selected benchmarks. This figure illustrates the historical investment performance of 4 different models with 6-month look-back period and 6-month holding period. In addition, three benchmark portfolios are shown – market index, 60–40 fixed mix portfolio and long-only momentum strategy. (**a**) Entire Sample Period: 1980–2007 (**b**) First Half of the Sample Period: 1980–1993 (**c**) Second Half of the Sample Period: 1994–2007

during the second sub-period, which corresponds to the period after the momentum effects become popular due to the publication of Jegadeesh and Titman (1993).

We are now ready to explain why the traditional Markowitz model has performed relatively well compared to its variants, especially when shorter holding periods are employed. The variants are designed to overcome the high sensitivity of the Markowitz model to input parameters. Therefore, in a sense, they smooth the result toward the direction that the optimal solutions wouldn't vary too much as the estimated input parameters change. So, higher weights on recent winners would be penalized, which would potentially discount the momentum effects. In contrast, the optimal weights from the Markowitz model is obtained via conventional optimization procedures, so it has a better chance of putting higher weights on the winners during the look-back period than its variants, and thus provides better utilization of the momentum effects. It gives us a simple, yet effective recipe to exploit the momentum effects in the context of the optimal asset allocation: investment performance can be improved by estimating the market parameters from relatively recent historical data (3 to 12 months).

18.5 Conclusions and Future Directions

In this paper, we construct various portfolios from four different optimal asset allocation models, and compare the historical performance for the period of time from 1980 to 2007. There are several meaningful findings: (1) the traditional Markowitz model has performed reasonably well as compared to its robust versions; (2) portfolios with shorter look-back periods (equal or less than a year) outperform ones with longer look-back periods in all cases; (3) these observations are in fact consistent with the momentum effects, which imply the recent winners tend to have better performance than the recent losers; and (4) investment performance can be improved by taking the momentum effects into account by utilizing the market parameter estimators from recent historical data.

What are possible extensions? First, it would be interesting to allow shorting of assets. Again, the momentum strategy could be the benchmark, since empirical results show that recent momentum losers will continue to underperform the market for subsequent periods. Shorting of industry-level assets is becoming more practical due to the emergence of the ETFs. In many cases, ETFs can be easily shorted.

Second, the described methodology can be applied to other extensions of the traditional Markowitz model, like Stein estimators (Jorion 1986). We suspect that the traditional Markowitz model will again perform relatively well.

The third domain for extending the analysis involves integrating asset management with borrowing (leverage) and other liability related issues – asset and liability management (ALM). In many of these applications such as pension plans, university endowments, and hedge funds, there are decided advantages to construct multi-period ALM models. The ex post success of momentum strategy shall apply to these optimization models. But this conjecture needs to be evaluated with real world experiences.

Appendix

Datastream industry classification

Level 2 (10 indices)	Level 3 (18 indices)	Level 4 (38 indices)
Oil and gas	Oil and gas	Oil and gas producers; oil equipment, services and distribution
Basic materials	Chemicals	Chemicals
	Basic resources	Forestry and paper; industrial metals; mining
Industrials	Construction and materials	Construction and materials
	Industrial goods and services	Aerospace and defense; general industrials; electronic and electrical equipment; industrial engineering; industrial teleportation; support services
Consumer goods	Automobiles and parts	Automobiles and parts
	Food and beverage	Beverages; food producers
	Personal and household goods	Household goods; leisure goods; personal goods; tobacco
Health care	Health care	Health care equipment and services; Pharmaceuticals and biotechnology
Consumer services	Retail	Food and drug retailers; general retailers
	Media	Media
	Travel and leisure	Travel and leisure
Telecommunication	Telecommunication	Fixed line telecommunication; mobile telecommunication
Utilities	Utilities	Electricity; gas, water and multi-utilities
Financials	Banks	Banks
	Insurance	Nonlife insurance; life insurance
	Financial services	Real estate; general financials; equity investment instruments
Technology	Technology	Software and computer services; technology hardware and equipment

Note: DataStream industry classification is almost identical to Dow-Jones/FTSE ICB (Industry Classification Benchmark).

References

Badrinath, S.G., Wahal, S., Momentum trading by institutions. Journal of Finance, 2002, 57.6, 2449–2478

Ben-Tal, A., Nemirovski, A., Lectures on modern convex optimization: Analysis, algorithms; engineering applications. SIAM-MPS Series in Optimization, 2001

Ben-Tal, A., Nemirovski, A., Optimal design of engineering structures. Optima, 1995, 47, 4–8

Black, F., Litterman, R., Asset allocation: Combining investor views with market equilibrium. Goldman, Sachs & Co., Fixed Income Research, 1990

Boyd, S., Vandenberghe, L., Convex optimization, Cambridge University Press, 2004

Carhart, M., On persistence in mutual fund performance. Journal of Finance, 1997, 52, 57–82

Cleary, S., Inglis, M., Momentum in Canadian stock returns. Revue Canadienne des Sciences de l'Administraition, 1998, 15.3, 279–291

Demir, I., Muthuswamy, J., Walter, T., Momentum returns in Australian equities: The influences of size, risk, liquidity and return computation. Pacific-Basin Finance Journal, 2004, 12, 143–158

El Ghaoui, L., Lebret, H., Robust solutions to least-squares problems with uncertain data. SIAM Journal of Matrix Analysis and Applications, 1997, 18.4, 1035–1064

Fabozzi, F.J., Kolm, P.N., Pachamanova, D., Focardi, S.M., Robust portfolio optimization and management, Wiley, 2007

Grauer, R.R., Hakansson, N.H., Returns on levered, actively managed long-run portfolios of stocks, Bonds and Bills, 1934–1983. Financial Analysts Journal, 1985, 41, 24–46

Grinblatt, M., Titman, S., Wermers, R., Momentum investment strategies, portfolio performance and herding: A study of mutual fund behavior. American Economic Review, 1995, 85, 1088–1105

Hameed, A., Kusnadi, Y., Momentum strategies: Evidence from Pacific Basin stock markets. Journal of Financial Research, 2002, 25.3, 383–397

Idzorek, T.M., A step-by-step guide to the Black-Litterman model. working paper, 2004

Jegadeesh, N., Titman, S., Returns to buying winners and selling losers: Implications for stock market efficiency. Journal of Finance, 1993, 48.1, 65–91

Jobson, J.D., Korkie, B.M., Putting Markowitz theory to work. Journal of Portfolio Management, 1981, 7 summer, 70–74

Jorion, P., Bayes-Stein estimation for portfolio analysis. Journal of Financial and Quantitative Analysis, 1986, 21.3, 279–292

Kang, J., Liu, M., Ni, S.X., Contrarian and momentum strategies in the China stock market: 1993–2000. Pacific-Basin Finance Journal, 2002, 10, 243–265

Larsen, G.A. Jr., Resnick, B.G., Parameter estimation techniques, optimization frequency and equity portfolio return enhancement. Journal of Portfolio Management, 2001, 27, 27–34

Liu, C., Lee, Y., Does the momentum strategy work universally? Evidences from the Japanese stock market. Asia-Pacific Financial Markets, 2001, 8, 321–339

Markowitz, H.M., Portfolio selection. Journal of Finance, 1952, 7.1, 77–91

Markowitz, H.M., The optimization of a quadratic function subject to linear constraints. Naval Research Logistics Quarterly, 1956, 3, 111–133

Meucci, A., Risk and asset allocation, Springer, New York, 2005

Michaud, R.O., Efficient asset management: A practical guide to stock portfolio optimization and asset allocation. Harvard Business School Press, 1998

Mulvey, J.M., Kim, W.C., Constructing a portfolio of industry-level momentum strategies across global equity markets. Working paper, Operations Research Financial Engineering Department, Princeton University, 2007

Mulvey, J.M., Kim, W.C., Active equity managers in the U.S.: Do the best follow momentum strategy? Journal of Portfolio Management, 2008a, 34 winter, 126–134

Mulvey, J.M., Kim, W.C., Improving equity diversification via industry-wide market segmentation. Journal of Financial Transformation, 2008b, Forthcoming

Mulvey, J.M., Simsek, K., Zhang, Z., Improving investment performance for pension plans. Journal of Asset Management, 2006, 7.2, 99–108

Nofsinger, J., Sias, R.W., Herding and feedback trading by institutional and individual investors. Journal of Finance, 1999, 54, 2263–2295

Pasotr, L., Portfolio selection and asset pricing models. Journal of Finance, 2000, 55.1, 179–223

Rouwenhorst, K.G., International momentum strategies. Journal of Finance, 1998, 53.1, 267–284

Sapp, T., Tiwari, A., Does stock return momentum explain the "smart money" effect? Journal of Finance, 2004, 59, 2605–2622

Satchell, S., Scowcroft, A., A demystification of the Black-Litterman model: Managing quantitative and traditional portfolio construction. Journal of Asset Management, 2000, 1.2, 138–150

Sias, R.W., Institutional herding. Review of Financial Studies, 2004, 17, 165–206

Sias, R.W., Starks, L.T., Titman, S., The price impact of institutional trading. Working Paper, 2001

Chapter 19
Reflections on Portfolio Insurance, Portfolio Theory, and Market Simulation with Harry Markowitz

Bruce I. Jacobs and Kenneth N. Levy

19.1 Introduction

We first became aware of Harry Markowitz's groundbreaking work in our early years in graduate school, where his notions of the efficient frontier, mean–variance optimization, and portfolio theory in general were required reading. When we founded Jacobs Levy Equity Management, we used our own quantitative methods [as originally discussed in Jacobs and Levy (1988)] to predict security returns, but we used Harry's ideas to form efficient portfolios from those security predictions.

We have since been fortunate to work with Harry Markowitz in several fields in which he was either the originator or was a major developmental force. We have found that one of the most intellectually satisfying experiences is to have a conversation with Harry that leads to a question or problem for which neither interlocutor has an answer. At this stage, Harry generally stops, sits back, and says, "Hmm... That's interesting." One then knows that something fascinating is about to happen. We hope that this essay provides the reader some insight into the way in which collaboration with Harry leads to new and interesting ideas.

Portfolio Insurance and Positive Feedback. Soon after we founded Jacobs Levy Equity Management, the market suffered the crash of October 1987. In the years leading up to the crash, portfolio insurance was very much in vogue, and we warned that such a positive feedback trading strategy had the potential to destabilize markets. We believed that portfolio insurance was a major cause of the crash, and we wrote a book (Jacobs 1999) on the topic. We sent a draft of the book to Harry, who not only liked it, but offered to write its foreword. In the foreword, in his subtle and piercing way, Harry makes the distinction between portfolio insurance and portfolio theory and their effects on financial markets.

Portfolio Theory. We knew that Harry was not only a portfolio theorist, but also was putting his theories into practice, managing a portfolio at Daiwa Securities. We discovered that Harry's expected return estimation procedures incorporated

B.I. Jacobs (✉) and K.N. Levy
Jacobs Levy Equity Management, Florham Park, New Jersey, USA
e-mail: bruce.jacobs@jlem.com

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_19,
© Springer Science+Business Media, LLC 2010

our ideas about disentangling the various sources of security price changes. Thus, Harry's portfolio management used some of our ideas, as we employed his in our portfolio optimization. As our relationship with Harry grew, we discovered other common areas of interest. In particular, we were interested in finding computationally efficient ways to compute optimal portfolios that included short positions. This interest led to the development of theorems regarding the conditions under which standard efficient algorithms could be applied to the long–short problem [as discussed in Jacobs et al. (2005)], and to the concept of trimability [as described in Jacobs et al. (2006)].

Market Simulation. One powerful weapon in Harry's formidable intellectual arsenal is his skill in computer simulation. Portfolio managers often forget that Harry not only created portfolio theory, but he is also a leading figure in the simulation world: he created the simulation language SimScript. We were very interested in studying the behavior of financial markets in response to various stimuli, but (for reasons we will discuss later) models that could simulate realistic markets were not available. Thus, we teamed up with Harry to create the Jacobs Levy Markowitz Simulator (JLMSim), as described in Jacobs et al. (2004).

Below, we briefly summarize our works with Harry, and some of our work that built on Harry's early work.

19.2 Portfolio Insurance and Positive Feedback

In Harry Markowitz's view, the causes of the crash of October 19, 1987 should be studied so that one can understand a tumultuous event in stock market history, and also so that one can grasp the implications for stock market mechanisms and their possible consequences. Harry believes, as we believe, that the severity of the 1987 crash was due, in large part, to the use of an option replication strategy known as "portfolio insurance." In the foreword to our book *Capital Ideas and Market Realities* (Jacobs 1999), Harry proposed a very simple thought experiment to compare the mean–variance efficiency of trading based on portfolio theory and that based on portfolio insurance.

Consider portfolios consisting only of a single security (the "market") and cash. Since portfolio insurance does not make use of beliefs about market movements, one can assume that the market's returns are independent and identically distributed. For simplicity, assume that the portfolio can be switched back and forth between cash and the market without cost.

To get some idea of the performance of portfolio insurance, compare a simple version of it to a constantly rebalanced portfolio. The portfolio insurance rule could be any function of past observations but, to obtain specific results, assume that over some number of time periods, the portfolio insurance rule directs the investor to be completely in cash for half the time and completely in the market for the other half. The alternative strategy is to rebalance the portfolio at each period to be half in the market and half in cash. Now compute the realized mean return and

Table 19.1 Mean and variance of return for two trading strategies

	$\mathbb{E}[r]$	$\mathbb{V}[r]$
Rebalanced 50/50 portfolio	$(r_m + r_0)/2$	$\sigma_m^2/4$
Switch back and forth	$(r_m + r_0)/2$	$\sigma_m^2/2 + (r_m - r_0)^2/4$

the variance of return for these strategies. Since trades executed by portfolio insurance strategies are not motivated by shifting beliefs about the market's movements, one can assume that such strategies' performance per period is a sequence of random draws.

With these simple assumptions, Harry computes the means and variances of the returns of the "switch back and forth" (i.e., portfolio insurance) strategy and the rebalanced strategy. The actual values are given in Table 19.1, in which $\mathbb{E}[r]$ is the mean of the strategies' return, $\mathbb{V}[r]$ is the variance of the returns, r_m is the mean value of the market's return, σ_m^2 is the variance of the market's return, and r_0 is the return on cash.

Table 19.1 shows that switching back and forth between cash and stocks is detrimental to mean–variance efficiency, *even assuming zero transaction costs*. These two strategies have the same mean return. However, the strategy of switching back and forth has more than twice the return variance of the strategy of rebalancing to a 50/50 portfolio. The rebalanced strategy will be on the market line, whereas the strategy of switching back and forth will be on the inefficient side of the market line. More generally, whatever the proportions of stock and cash chosen, given the assumption of independent, identically distributed returns, the rebalanced portfolio will lie on the market line, whereas the portfolio that switches back and forth will lie on the inefficient side of the market line. Even if one uses semivariance as a measure of risk, as discussed in Markowitz (1959, Chap. 2), the rebalanced portfolio will be on the (semivariance) market line, whereas the strategy that switches back and forth will be on the inefficient side of the line.

The given example is, of course, an extreme case in which the portfolio insurer is either in stocks or in cash, but not both simultaneously. The direction of the result is the same, however, in the more realistic, less extreme, case where the proportion of stocks held by the portfolio insurer varies over time as a function of past observations. With his characteristic wit, Harry provided a fictitious debate in which a portfolio *insurance* supporter sparred with a portfolio *theory* supporter over the given example.

In the debate, the portfolio insurance supporter argued that the strategy's mean–variance inefficiency is the price paid to reshape the probability distribution of returns over a longer interval of time; i.e., the insured period. The portfolio insurance supporter pointed out that, under reasonable assumptions, the greatest loss in any period is less for the portfolio insurance strategy than for the rebalanced strategy and, given the assumptions, in no month does the loss exceed the preset floor.

In response, the portfolio theory supporter countered that, for the period of analysis as a whole, the rebalanced portfolio grew more than the switched-back-and-forth one. This is because the strategy with the greater average logarithmic return[1] will have grown the most during the period; average logarithmic return is very closely approximated by a function of return mean and variance [see Markowitz (1959, Chap. 6)]; and this approximate average decreases with increasing variance. A particular point on the mean–variance frontier gives approximately maximum growth. Moreover, every point on the frontier gives approximately maximum growth in the long run for given short-run fluctuations.

In rebuttal, the portfolio insurance supporter countered that, after one or two bad years, an investor (the client) hiring an investment manager (the agent) using the rebalancing strategy might not wait for the long run, but might summarily fire the investment manager. To which, the portfolio theory supporter might contrast the needs of the client with those of the agent, thereby seeking the moral high ground. The portfolio insurance supporter might state that, in practice, applications of portfolio theory sometimes put investment manager motives ahead of true client needs, as perhaps when mean–variance analysis is used for a given average return to minimize tracking error rather than total variability. Finally, the portfolio theory supporter could deliver the winning argument by stating that, whatever the arguments pro and con, the debate was irrelevant because portfolio insurance simply did not work in practice, especially when it was needed the most.

Given that it is meaningless to argue about whether portfolio insurance would have been the right thing to do had it worked, it is still instructive to study its effects. Portfolio insurance destabilized the market, creating liquidity problems that effectively caused it to fail. Those following portfolio insurance strategies bought securities when the market went up and sold when the market went down. Such a strategy creates positive feedback, reinforcing upward market moves and exacerbating downturns. Destabilization is a well-known consequence of positive feedback. In contrast, the rebalancing strategies that portfolio theory implies when used to maximize return sell securities when the market rises and buy when the market falls.[2] This is negative feedback, and it tends to stabilize the market. Thus, in Harry's words, such an application of portfolio theory is, if nothing else, more environmentally friendly than portfolio insurance.

Jacobs (2004, 2009) examines more recent episodes of market instability attributable to positive feedback strategies. In particular, in the middle of this decade structured finance instruments, including residential mortgage-backed securities and collateralized debt obligations as well as a type of credit insurance known as credit default swaps, facilitated a positive feedback system. The ability to transfer risk from lenders to investors and insurers encouraged mortgage lending and lowered mortgage rates, enabling a housing bubble to develop. Increasing home prices in turn

[1] By logarithmic return, we mean $\log((p_{t+1} + \mathrm{d}_t)/p_t)$, where p_t is the price at instant t and d_t is the dividend paid immediately after that instant.

[2] Investors often establish a policy portfolio consisting of an unleveraged mix of equity, fixed income, and other assets. If equities rise in price, the percentage of the portfolio held in equities will exceed the policy portfolio's percentage, and equities will be sold (absent a change in the investor's beliefs) to rebalance the portfolio.

increased demand for mortgages and creation of more structured finance products and credit default swaps. As the experience with portfolio insurance should have taught us, however, shifting risk may change *who* the risk holders are, but it does not reduce overall risk and can actually increase it. When housing prices leveled out and started to decline in 2006, the same products that had been used to (supposedly) reduce risk ended up spreading risk through the entire financial system.

Instruments and strategies that purport to reduce systematic risk for portfolios, including stock and mortgage portfolios, can end up increasing risk for the overall financial system. Risk bearers need to be able to withstand unexpected losses; otherwise the risk can become systemic and, as Jacobs (2004) warned, fall on taxpayers. In 2008–2009, the risk bearers, including financial institutions that bought structured products and underwrote credit default swaps, failed. The government – that is, the taxpayer – had to step in as the risk bearer of last resort.

19.3 Portfolio Theory

We were delighted when Harry wrote the foreword to our second book, *Equity Management: Quantitative Analysis for Stock Selection* (Jacobs and Levy 2000). Harry noted in the foreword, "It may be fairly asserted that Jacobs and Levy's work is based on mine, and my work is based on theirs." He points out that we, as do almost all quantitative investment firms, make use of the general mean–variance portfolio selection model presented in Markowitz (1956, 1959), which in turn extended a proposal in Markowitz (1952). This is the sense in which some of our work is based on his.

To be practically applicable, mean–variance analysis as presented in Markowitz (1952, 1956, 1959) requires estimates of the means and variances of the returns of individual securities, as well as covariances between returns of pairs of securities. But those pioneering articles did not specify how to make these estimates. In fact, Markowitz (1952) begins: "The process of selecting a portfolio may be divided into two stages. The first stage starts with observation and experience and ends with beliefs about the future performances of available securities. The second stage starts with the relevant beliefs about future performances and ends with the choice of portfolio. This paper is concerned with the second stage." It turned out that when Harry addressed the first stage, and turned his hand to portfolio management, he and his colleagues used expected return estimation procedures based on Jacobs and Levy (1988a), as cited in Bloch et al. (1993). We had devised a multivariate approach to return estimation that took into account a multitude of factors and their interrelationships (Jacobs and Levy 1988a). Before 1988, researchers had generally examined only one to three variables at a time. We fit a series of monthly cross-sectional regressions of security excess returns against 25 factors, including analysts' earnings estimates, earnings surprises, cash flow/price value, and tax-loss selling, together with industry affiliations. We were thus able to disentangle "pure," statistically significant pricing inefficiencies from factors that were merely "naïve" surrogates for other factors. We also examined (Jacobs and Levy 1988b) how abnormal equity returns were associated with the turn of the year, the week, and the month, as well as with holidays and time of day, and how payoffs to the size effect are predictable using macroeconomic drivers such as interest rates and industrial production (Jacobs and Levy 1989). Harry used some of these ideas in his return estimation.

This section describes some of the portfolio theory work that we have done with Harry, or that we have based on Harry's work. We start in Sect. 19.3.1 with a description of our work on integrated optimization and its relationship to Harry's "Rational Decision Maker." Then, in Sect. 19.3.2, we describe the work we did together on the topic of trimability. In Sects. 19.3.3 and 19.3.4, we discuss some equivalences discovered.

19.3.1 Integrated Optimization and Unnecessary Constraints

To maximize portfolio mean return for given portfolio return variance, or minimize portfolio return variance for given portfolio mean return, one should not impose any constraints that are not absolutely required. In Jacobs et al. (1998), we define a minimally constrained portfolio that maximizes expected investor utility and argue that imposing any other constraints can only reduce utility (or, at best, not increase it). In that paper, we also define and advocate the use of "integrated optimization."

In his foreword to Jacobs and Levy (2000), Harry notes that our work on integrated optimization of long–short portfolios and the estimation of security expected returns is "to be acknowledged for bridging the gap between theory and practice in the world of money management." He goes on to say that the translation of investment ideas into products and strategies must involve trade-offs between theory and practice. He then discusses why, in the portfolio optimization problem, investors might want to add constraints on position sizes and sectors, despite the theoretical cost of these constraints.

As Harry explains with reference to Markowitz (1959, Chap. 13), the mean–variance investor approximates a rational decision maker (RDM) acting under uncertainty. The mean–variance optimal portfolio may be less averse to an extreme downside move than the one that optimizes an investor's true (i.e., subjective) expected utility [see Table 1 in Levy and Markowitz (1979)]. It is therefore possible that adding constraints to a minimally constrained mean–variance analysis may produce a portfolio that gives higher true expected utility, even though it gives a lower value to a mean–variance approximation.

Nevertheless, as we pointed out in Jacobs et al. (1998), a general principle of optimization is that constrained solutions do not offer the same level of utility as unconstrained solutions unless, by some fortunate coincidence, the optimum lies within the feasible region dictated by the constraints.

Treynor and Black (1973, page 66) had hinted at similar issues, and specifically posed the following question: "Where practical is it desirable to so balance a portfolio between long positions in securities considered underpriced and short positions in securities considered overpriced that market risk is completely eliminated?" We reformulated Treynor and Black's question slightly, posing the following three questions:

1. Under what conditions will a net holding of zero (i.e., dollar-neutrality) be optimal for a long–short portfolio?
2. Under what conditions will the combined optimal holdings in a long–short portfolio be beta-neutral?
3. Under what conditions will dollar-neutrality or beta-neutrality be optimal for the active portion of an equitized long–short portfolio?

To answer these questions, consider the standard mean–variance utility function:

$$U = E_P - \frac{1}{2\tau} V_P, \tag{19.1}$$

where E_P is the expected return on the investor's portfolio, $V_P = \sigma_p^2$ is the variance of the return, and τ is the investor's risk tolerance.

Assume that, in seeking to maximize the utility function in (19.1), the investor has an available capital of K dollars and has acquired n_i shares of security $i \in \{1, 2, \cdots, N\}$. A long holding is represented by a positive number of shares, and a short holding is represented by a negative number. The holding h_i in security i is the ratio of the amount invested in that security to the investor's total capital. Thus, if security i has price p_i, then $h_i = n_i p_i / K$. With these definitions, the portfolio's return has mean and variance given by

$$E_P = h^\top r, \tag{19.2}$$

$$\sigma_P^2 = h^\top Q h, \tag{19.3}$$

where h is a vector containing the individual holdings, r is the vector of expected security returns, and Q is the covariance matrix of the securities' returns.[3] Using (19.2) and (19.3), one finds that the *unconstrained* portfolio vector that maximizes the utility in (19.1) is

$$h = \tau Q^{-1} r. \tag{19.4}$$

This unconstrained portfolio will be naturally dollar-neutral (i.e., dollar-neutral without the need to impose any constraints) if the net holding, H, is zero. Using a constant correlation model as described in Elton et al. (1976), we found that this net holding is closely approximated by

$$H = \frac{\tau}{1-\rho} \sum_{i=1}^{N} (\xi_i - \bar{\xi}) \frac{r_i}{\sigma_i}, \tag{19.5}$$

[3] Similar expressions are obtained whether one works with absolute or excess returns.

where ρ is the correlation from the constant correlation model, and, for security i, r_i is the expected return, σ_i is the standard deviation of the return, $\xi_i = 1/\sigma_i$ is a measure of the return stability, and $\bar{\xi}$ is the average of all the ξ_i.

The net holding in (19.5) will be zero either in the trivial case when the risk tolerance is zero, or in the more interesting case when the sum is zero. The sum can be regarded as the net risk-adjusted return (r_i/σ_i) of all securities weighted by the deviation $(\xi_i - \bar{\xi})$ of their stability from the average stability.[4] If this sum is positive, the net holding should be long. Conversely, if this sum is negative, the net holding should be short.

The fractional change in utility when dollar-neutrality is imposed is

$$\frac{\Delta U}{U} = -\frac{(\mathbf{1} + Q^{-1}r)^2}{(\mathbf{1}^\top Q^{-1}\mathbf{1})(r^\top Q^{-1}r)},$$

where $\mathbf{1}$ is a vector of ones. This change has a maximum value of zero (which occurs when the condition for dollar-neutrality is satisfied) and is otherwise always negative. Thus, only under the special condition in which H in (19.5) is equal to zero will the optimal portfolio be dollar-neutral. Constraining the holding to be zero when this condition is not satisfied will produce a suboptimal portfolio; i.e., one with decreased (mean–variance) utility.

To answer the second question about beta-neutrality, we performed a similar analysis to that above, using Sharpe's single index model. In this case, we found that the beta of an unconstrained portfolio is approximately

$$\beta_P = \frac{\tau}{1-\rho} \sum_{i=1}^{N} \left(\beta_i - \bar{\beta}\right) \frac{r_i}{\sigma_i}, \tag{19.6}$$

where

$$\bar{\beta} = \frac{1}{N} \sum_{i=1}^{N} \frac{\beta_i}{\sigma_i}.$$

Equation (19.6) is entirely analogous to (19.5): the sum can be regarded as the net risk-adjusted return (r_i/σ_i) of all securities weighted by the deviation $(\beta_i - \bar{\beta})$ of their beta from the volatility-weighted beta. The net beta should have the same sign as this sum. Only under the special condition in which β_p in (19.6) is equal to zero will the optimal portfolio be beta-neutral. Constraining the beta to be zero when this condition is not satisfied will produce a suboptimal portfolio. Thus we concluded that only under very specific conditions will dollar- or beta-neutrality be optimal.

The same conclusions hold with regard to the third question. That is, only under very specific conditions will an equitized long–short portfolio hold long and short positions that are balanced by dollar or beta. Furthermore, the degree of equitiza-

[4] Note that the sum is not a weighted average because the weights do not sum to 100%, and some weights are in fact negative.

tion itself becomes a matter of optimization. As we state in Jacobs et al. (1998, page 40), "The important question is not how one should allocate capital between a long-only portfolio and a long–short portfolio but, rather, how one should blend active positions (long and short) with a benchmark security in an integrated optimization." Jacobs et al. (1999) show that a theoretically optimal portfolio would be constructed in a single, integrated optimization that considers the expected returns, risks, and correlations of all securities, including any benchmark, simultaneously. Such a portfolio will rarely be naturally totally neutral with respect to any particular characteristic.

Of course, there may be perfectly valid tax, accounting, or regulatory reasons for dollar-neutral, beta-neutral, market-neutral, or fully equitized portfolios. Such portfolios may also be preferred for behavioral reasons, such as mental accounting, or because they fit more readily into established frameworks for performance evaluation and comparison. But perhaps such constrained portfolios merely reflect, as Harry argues, "the inability of human decision makers to fully emulate [purely] R[ational] D[ecision] M[akers] in maximizing expected utility in the face of uncertainty and illiquidity."

The general theoretical conclusion, however, is that imposing neutrality moves the portfolio away from mean–variance optimality. The corollary to this finding is that determining equity market exposure should be done as part of determining individual security positions: active long and short positions, as well as benchmark holdings, should be determined jointly, in an integrated optimization. Harry's 1952 tenet still holds: mean–variance analysis provides the right kind of diversification for the right reason. However, imposing unnecessary constraints can cause a portfolio to be mean–variance suboptimal.

19.3.2 Trimability

Following our work on the optimality of long–short portfolios and the benefits of integrated optimization, we turned our attention to fast methods for optimizing such long–short portfolios subject to realistic constraints. Harry's expertise in optimization theory, and particularly in the area of optimizing quadratic functions subject to linear constraints (see Markowitz 1956), proved invaluable. This type of optimization is directly applicable to optimization of long–short portfolios.

The portfolio optimization is framed as a quadratic function optimization as follows: Consider a portfolio consisting of n securities with expected returns $\mu_1, \mu_2, \cdots, \mu_n$. The portfolio can include both risky and riskless securities. The portfolio's expected return, E_P, is a weighted sum of the n security returns:

$$E_P = \sum_{i=1}^{n} x_i \mu_i , \tag{19.7}$$

where $x_1, x_2, \cdots, x_n$ are the security weights in the portfolio. If the covariance between the returns of security i and security j is σ_{ij}, the portfolio's return variance, V_P is

$$V_P = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i \sigma_{ij} x_j. \tag{19.8}$$

The security weights may be subject to various constraints. For long-only portfolios, common constraints include the following:

$$\sum_{k=1}^{n} a_{jk} x_k = b_j, \quad \text{for } j = 1, \cdots, m \tag{19.9}$$

and

$$x_i \geq 0, \quad \text{for } i = 1, \cdots, n, \tag{19.10}$$

where m is the number of constraints. Equation (19.9) might include, for example, a budget constraint according to which the sum of the weights must equal a fixed number. Equation (19.10) is a nonnegativity constraint.

The general single-period mean–variance portfolio selection problem is, for all variances, V_P, find the corresponding portfolios that provide maximum expected return E_P, or, alternatively, for all expected returns E_P, find the corresponding portfolios that provide minimum variance V_P, subject to the given constraints.

For a long–short portfolio, the sign of x_i is not constrained. A negative value of x_i is interpreted as a short position. Unfortunately, with such an interpretation, unrealistic portfolios can be obtained. For example, if as in the Capital Asset Pricing Model (CAPM), the portfolio is subject only to the full investment constraint, an investor could deposit \$1,000 with a broker, short \$1,000,000 of Stock A, and use the proceeds plus the original deposit to purchase \$1,001,000 of Stock B. Short positions do not, in fact, work this way.

Although no single constraint set applies to all long–short portfolios, all constraints of practical interest can be accommodated if one adopts the convention of representing an n-security long–short portfolio in terms of $2n$ nonnegative variables, $x_1, \cdots, x_{2n}$ in which the first n variables represent the securities in a given set held long, the second n variables represent short sales in the same set of securities, and one chooses the long–short portfolio subject to the following constraints:[5]

$$\sum_{k=1}^{2n} a_{jk} x_k = b_j, \quad \text{for } j = 1, \cdots, m \tag{19.11}$$

and

$$x_i \geq 0, \quad \text{for } i = 1, \cdots, 2n. \tag{19.12}$$

[5] Because the second n variables represent short sales in the same set of securities, if security i is held long, x_i will be positive, and if security i is sold short, x_{n+i} will be positive.

The types of constraints incorporated in (19.11) and (19.12) include budget constraints, upper and lower bounds on long and short positions, equality constraints on particular positions, market-neutrality constraints, constraints on net long or short positions, or on borrowing or margins. An apparent disadvantage of (19.11) and (19.12), insofar as portfolio optimization is concerned, is that they allow long and short positions in the same security. We consider this issue later.

In general, the covariances, σ_{ij}, in (19.8) are nonzero, so the covariance matrix will be dense (i.e., will contain mostly nonzero entries). The solution of the general mean–variance portfolio selection problem requires the inversion of a matrix that includes this covariance matrix as one of its blocks. This inversion is one of the major computational burdens in portfolio optimization.

It was unclear whether fast portfolio optimization algorithms, which were applicable to long-only portfolios, were applicable to long–short portfolios as well. Long–short portfolios can take many forms, including market-neutral equity portfolios that have a zero market exposure and enhanced active equity portfolios that have a full market exposure, such as 120–20 portfolios (with 120% of capital long and 20% short). While studying the problem of optimizing long–short portfolios with Harry, we collectively came up with the notion of "trimability." This is a sufficient condition under which a fast portfolio optimization algorithm designed for long-only portfolios will find the correct long–short portfolio, even if the algorithm's use would violate certain assumptions made in the formulation of the long-only problem.[6] In the following, we briefly describe the basic approach of using covariance models to design fast portfolio optimization algorithms, and then discuss the trimability condition under which such algorithms are also applicable to long–short portfolios.

For long-only portfolios, there are at least three types of models – factor models, scenario models, and historical models – that can be used to transform the portfolio selection problem into one that requires the inversion of a diagonal (or nearly diagonal) matrix. Diagonal matrices are easy to invert, so their use in place of denser matrices can greatly simplify and speed the optimization problem. The "trick" to obtaining diagonal matrices for long-only portfolios is to introduce fictitious securities that are linearly related to the original securities but constrained in some way. For example, consider a factor model in which r_i, the return of security i, is given by

$$r_i = \alpha_i + \sum_{k=1}^{K} \beta_{ik} f_k + u_i, \quad \text{for } i = 1, \cdots, n, \tag{19.13}$$

where α_i is a constant, f_k is the return on the k-th common factor, β_{ik} is the factor loading, K is the number of common factors, and u_i is an idiosyncratic term

[6] The mathematical specifics of this condition are described in detail in Jacobs et al. (2005).

assumed to be uncorrelated with u_j for all $i \neq j$ and uncorrelated with all f_k for $k = 1, \cdots, K$. For simplicity, we also assume that f_k is uncorrelated with f_j for $j \neq k$.[7]

To perform the diagonalization, one introduces fictitious securities, one for each common factor (see Sharpe 1963; Cohen and Pogue 1967), with the weight of each fictitious security constrained to be a linear combination of the weights of the real securities. Accordingly, one defines a set of K fictitious securities with weights $y_1, \cdots, y_K$ in terms of the real securities as follows:

$$y_k = \sum_{j=1}^{n} x_j \beta_{jk}, \quad \text{for } k = 1, \cdots, K. \tag{19.14}$$

With this definition, the portfolio variance can be written (see Jacobs et al. 2005) in the form

$$V_P = \sum_{i=1}^{n} x_1^2 V_i + \sum_{k=1}^{K} y_k^2 W_k, \tag{19.15}$$

where W_k is the variance of f_k. Equation (19.15) expresses V_P as a positively weighted sum of squares in the n original securities and K new fictitious securities, which are linearly related to the original securities by (19.14).

Note that the variance expression in (19.15) contains only two single sums (whereas the variance expression in (19.8) contained a nested double sum). Therefore, (19.15) can be written in terms of a diagonal covariance matrix; i.e., we have effectively diagonalized the model.

We showed in Jacobs et al. (2005, 2006) that analogous procedures can be used to write scenario models and historical models in diagonal form. We refer to these models as "diagonalizable models." In each case, diagonalization transforms the variance expressions from ones couched in terms of dense covariance matrices to ones containing matrices that are slightly larger but have nonzero entries only along their diagonals. Inversion of such matrices is trivial.

Can this diagonalization procedure, used for long-only portfolios, be applied to the optimization of long–short portfolios? To investigate this question, we adopt the convention of representing an n-security long–short portfolio in terms of $2n$ nonnegative variables $x_1, \cdots, x_{2n}$. Let r_c be the return on cash or collateral. The portfolio's return R_P, is then

$$R_P = \sum_{i=1}^{n} r_i x_i + \sum_{i=n+1}^{2n} (-r_{i-n}) x_i + r_c \sum_{i=n+1}^{2n} h_{i-n} x_i. \tag{19.16}$$

[7] The mathematical details of the more general case, in which the factors are not necessarily mutually uncorrelated, are discussed in Jacobs et al. (2005).

The first term on the right of (19.16) represents the return contribution of the securities held long. The second term represents the contribution of the securities sold short. The third term represents the short rebate, where

$$h_i \leqslant 1, \quad \text{for } i = 1, \cdots, n$$

is the investor's portion of the interest received on the proceeds from the short sale of security i. With these definitions, the returns on the short positions are

$$r_i = h_{i-n} r_c - r_{i-n}, \quad \text{for } i = n+1, \cdots, 2n. \tag{19.17}$$

Let μ_i be the expected value of r_i, for $i = 1, \cdots, 2n$. Then, the expected return of the long–short portfolio is

$$E_P = \mathbb{E}[R_P] = \sum_{i=1}^{2n} x_i \mu_i. \tag{19.18}$$

To diagonalize, we assume a multifactor model with returns given by (19.13) and we define K new fictitious securities, $y_1, \cdots, y_K$, in terms of the real securities, as follows:

$$y_k = \sum_{j=1}^{n} x_j \beta_{jk} - \sum_{j=1}^{n} x_{n+j} \beta_{jk}, \quad \text{for } k = 1, \cdots, K.$$

From this definition, it follows (see Jacobs et al. 2005) that the variance of the portfolio's return is

$$V_P = \sum_{j=1}^{2n} x_i^2 V_i + \sum_{k=1}^{K} y_k^2 W_k - 2 \sum_{i=1}^{n} x_i x_{n+i} V_i. \tag{19.19}$$

Equation (19.19) is the expression for the variance of the return of a long–short portfolio when a multifactor covariance model is assumed. Note that, with the exception of the cross-product terms, (19.19) has exactly the same form as (19.15). Had the cross-product terms $x_i x_{n+i}$ been zero, the model for the long–short portfolio would have been diagonal.

Recall that x_i is the magnitude of a long position in security i, and x_{n+i} is the magnitude of a short position in security i. Therefore, if the cross-products are all zero, the portfolio has no simultaneous long and short positions in the same securities because either x_i or x_{n+i} is zero, or both are zero. We refer to such a long–short portfolio as a "trim" portfolio. Mathematically, a trim portfolio has

$$x_i x_{n+i} = 0, \quad \text{for } i = 1, \cdots, n.$$

Trim portfolios have the useful property that, for them, (19.19) has precisely the same form as (19.15); i.e., their covariance matrices (including fictitious securities) are diagonal.

Conceptually,[8] if, using the given return model, we are able to transform a feasible portfolio that is untrim (i.e., one that has at least one security in which it has simultaneous long and short positions) into a feasible portfolio that is trim in a way that does not reduce the portfolio's expected return, the model satisfies the "trimability condition."[9] Such a model is called "trimable." Importantly, we can apply existing fast portfolio optimization algorithms to trimable long–short portfolio models.

For a guarantee that an efficient set for a model in which the cross-product terms are ignored is an efficient set for the model in which they are not, we must be able to trim the model in the following way:

- remove the overlap from simultaneous long and short positions in each security in such a way that the smaller of the two positions diminishes to zero,
- add the overlap to a risk-free security holding,
- leave all other risky security holdings unchanged,
- maintain feasibility, and
- not reduce the expected return of the portfolio.

Although models with arbitrary constraint sets may not satisfy the trimability condition, a wide variety of constraints met in practice do satisfy it. In Jacobs et al. (2005, 2006), we provide examples of models that can be trimmed, as well as examples that cannot. We also provide tables that show the dramatic improvement in computational speed that can be achieved using fast algorithms to optimize trimable long–short portfolios.

19.3.3 Trim Equitized and Enhanced Active Equity Equivalence

Jacobs and Levy (2007) applies the concept of trimability to illustrate the relationship between equitized market-neutral long–short (ELS) portfolios and enhanced active equity (EAE) portfolios, such as 130–30 portfolios, and to show specifically that every ELS portfolio has an equivalent EAE portfolio, and vice versa.

In EAE portfolios, the strict long-only constraint is relaxed so that the manager can sell stocks short up to some prespecified percentage of capital (e.g., 30%), and use the proceeds of the short sales to buy additional long positions. The overall portfolio thus has 130% of its capital long and 30% short. Overall, it maintains a 100% exposure to the market.

An ELS portfolio also has a 100% exposure to the market, achieved with stock index futures or exchange traded funds (ETFs), and it has a long–short component that may have 100% of capital long, and 100% of capital short.

The EAE portfolio is essentially a compact form of the ELS portfolio. If the ELS portfolio contains short positions in stocks that are held in the equitizing instrument

[8] More precise statements are provided in Jacobs et al. (2005, 2006).

[9] This condition is called "Property P" in Jacobs et al. (2005).

(i.e., in the underlying index of the stock index future, or in the ETF), then the ELS portfolio is untrim. While the ELS portfolio may not be trimable in practice because individual securities in the equitizing instrument cannot be sold to remove overlaps, there is a unique EAE portfolio that is functionally identical to, but more compact than, the untrimmed ELS portfolio.[10]

Consider a market-neutral long–short portfolio that has $100M\%$ of capital long and $100M\%$ short, where M is a multiple of the investor's capital.[11] An equitized portfolio consisting of this market-neutral long–short portfolio and a benchmark index overlay is equivalent to an enhanced active equity portfolio with $100(1+E)\%$ held long and $100E\%$ sold short. Here, E is a quantity that we call the enhancement, equal to:

$$E = M - T,$$

where

$$T = \sum_{i \in S} \min\{|x_i|, b_i\}$$

is the fraction of capital trimmed to eliminate simultaneous long and short exposures to the same security, x_i is the weight of the i-th security in the market-neutral long–short portfolio, b_i is its weight in the benchmark, and S is the set of securities sold short in the market-neutral long–short portfolio. The trimmed amount, T, has a minimum value of zero (corresponding to the case where there is no overlap) and a maximum value of one (corresponding to the case where there is complete overlap). More details, including a comparison of EAE portfolios to ELS portfolios, and examples of equivalent portfolios, are provided in Jacobs and Levy (2007).

19.3.4 Algorithmic Equivalence

In considering fast algorithms, we were struck by the fact that some algorithms appeared to take completely different approaches, yet must produce the same efficient frontier. In particular, under realistic assumptions, there is a piecewise linear set of portfolios that supplies one and only one efficient portfolio for each efficient risk-return combination. If the covariance matrix is nonsingular, then this set of efficient portfolios is unique. Thus, any algorithm that traces out the mean–variance efficient frontier must produce the same results. One such algorithm is the critical line algorithm (CLA) [described in detail in Markowitz (1956, 1959, 1987)], which is an iterative technique, applicable to the general portfolio problem. Sharpe (1963) presented a procedure that greatly simplifies the CLA computation, specifically for a one-factor model of covariance with long positions only. Elton et al. (1978) presented alternative algorithms for finding the efficient frontier for various

[10] For transaction cost differences between EAE and ELS portfolios, see Jacobs and Levy (2007).

[11] When $M = 1$, the portfolio is a fully invested market-neutral long-short portfolio, with 100% of capital long and 100% of capital short.

special models, including the one-factor model with long positions only. Though they must produce the same efficient frontier, the two algorithms are parameterized differently.

Both the CLA and the Elton–Gruber–Padberg (EGP) algorithm trace out the same unique efficient frontier by varying a parameter in discrete steps over a certain range, and finding the corner portfolio that corresponds to the value of the parameter at each step. The CLA is applicable to arbitrary covariance models, while the EGP algorithm applies only to certain specific models of covariance. The uniqueness of the efficient frontier guarantees that the two algorithms must both find the same set of corner portfolios. Therefore, there must exist a unique relationship between the parameters used in the algorithms. In Jacobs et al. (2007), we explain that relationship for long-only portfolios with the assumption that the investor can neither lend nor borrow at the risk-free rate.

The relationship between E_P and V_P along the efficient frontier is shown in Fig. 19.1. This figure draws expected return on the horizontal axis, and variance on the vertical axis, as done in Markowitz (1987). This differs from the current convention of drawing standard deviation on the horizontal axis and expected return on the vertical axis. The broad line in the figure represents the efficient frontier itself. The parameter λ_E can be interpreted as half the slope of the efficient frontier in $(E_P,\ V_P)$ space; or, since[12] $\lambda_E > 0$,

$$\frac{\mathrm{d}E_P}{\mathrm{d}V_P} = \frac{1}{2\lambda_E}. \tag{19.20}$$

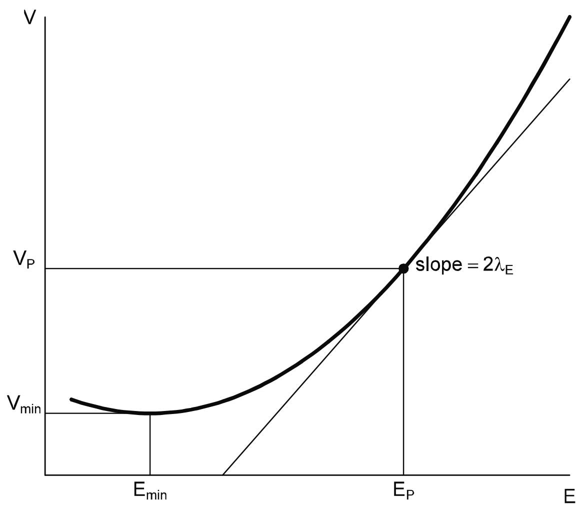


Fig. 19.1 Geometry of the CLA

[12] See Markowitz (Chap. 7, 1987) for a discussion of the CLA for the step in which λ_E actually reaches zero.

Each portfolio corresponds to a point at which a line with slope $2\lambda_E$ is tangent to the efficient frontier. As the slope of this line varies from infinity down to zero, the tangent point traces out the entire efficient frontier from its high-return, high-risk extreme down to $(E_{\min},\ V_{\min})$.

Unlike the CLA, which finds all corner portfolios by varying the slope term λ_E, the algorithm of Elton et al. (1978) finds all corner portfolios by varying an intercept term R_f. In the absence of a risk-free security, Elton, Gruber, and Padberg define a parameter

$$\lambda = R_f - R_f^0,$$

where R_f^0 is an intercept term for which the algorithm has already determined an optimal portfolio; i.e., it is the intercept term corresponding to the previous corner portfolio.

The relationship between $\sigma = \sqrt{V_p}$ and E_P along the efficient frontier is shown in Fig. 19.2. The broad line in the figure represents the efficient frontier itself. The EGP algorithm traces out the efficient frontier by finding each tangency point corresponding to a particular value of the intercept term R_f as it increases from some preset minimum value up to a preset maximum value. The relationship between the parameter R_f and the corresponding tangency portfolio is illustrated in Fig. 19.2. From the geometry of Fig. 19.2, it must be true for any (E_P, σ_P) pair along the efficient frontier that

$$\frac{\mathrm{d}E_P}{\mathrm{d}\sigma_P} = \frac{E_P - R_f}{\sigma_P}. \tag{19.21}$$

Now, since $V_P = \sigma_P^2$, where σ_P is the standard deviation of the portfolio's return, we have

$$\frac{\mathrm{d}E_P}{\mathrm{d}V_P} = \frac{\mathrm{d}E_P}{\mathrm{d}\sigma_P}\frac{\mathrm{d}\sigma_P}{\mathrm{d}V_P} = \frac{\mathrm{d}E_P}{\mathrm{d}\sigma_P}\frac{1}{2\sigma_P}.$$

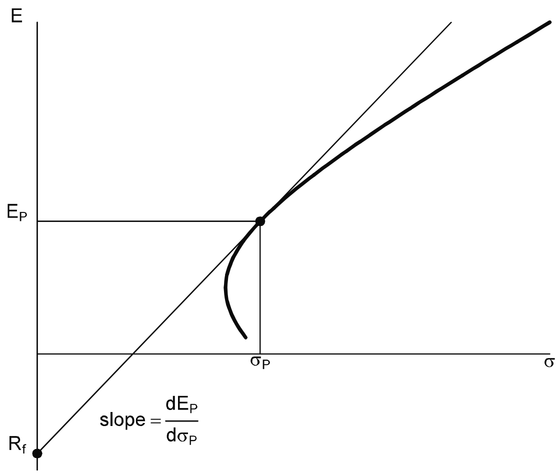


Fig. 19.2 Geometry of the EGP algorithm

Using (19.21), this becomes

$$\frac{\mathrm{d}E_P}{\mathrm{d}V_P} = \frac{E_P - R_f}{\sigma_P}\frac{1}{2\sigma_P} = \frac{E_P - R_f}{2V_P}. \tag{19.22}$$

Equating the derivatives in (19.20) and (19.22), we obtain

$$\frac{1}{\lambda_E} = \frac{E_P - R_f}{V_P}.$$

This is true for any (E_P, V_P), pair along the efficient frontier. Therefore, in particular, it must be true for the pair $(E_{\min}, V_{\min})$, so we find that

$$R_f = E_{\min} - \frac{V_{\min}}{\lambda_E},$$

showing that a constant relationship exists between R_f (the parameter varied in the EGP algorithm) and λ_E (the parameter varied in the CLA). Thus, we have unified the CLA and the EGP algorithm.

19.4 Market Simulation

One of our longer-term initiatives with Harry was to design and build a simulator that could explain the behavior of the market better than current models. Many market models use continuous-time methods (such as that used in the Black–Scholes–Merton option-pricing model that formed the basis of portfolio insurance). These models may use assumptions – for example, that the underlying security price process is fixed and that prices change randomly and continuously over time. The models are often useful because they can be solved analytically. They are not useful, however, when investment actions or changes in the underlying environment alter the price process. Nor can they tell us whether microtheories about the behavior of investors can explain the observed macrophenomena of the market.

We developed a market simulator, the Jacobs Levy Markowitz Simulator (or JLMSim), that has the potential to address these problems. JLMSim is an asynchronous-time simulator. It assumes that price changes reflect events, which can unfold in an irregular fashion. The price process of securities is not fixed, but is the result of simulated market participants trading with one another to maximize their own individual utility functions as conditions change and as random money flows occur into or out of the market. JLMSim allows users to model financial markets using their own inputs about the numbers and types of investors, traders, securities, and other entities that would have a bearing on markets in the real world.

Asynchronous models such as that used in JLMSim may also be better than continuous-time models for analyzing whether microtheories about investor behavior can explain market macrophenomena. From time to time, the market manifests liquidity black holes, which seem to defy rational investor behavior. One extreme case was the stock market crash on October 19, 1987. When prices fell precipitously and discontinuously on that day, rational value investors should have stepped in to pick up bargain stocks, but few did. Asynchronous models are able to explain both the abundance of sellers and the dearth of buyers. Our experiments with JLMSim show that only a relatively small proportion of momentum investors can destabilize markets, overwhelming value investors. Similarly explosive behavior can result when traders do not anchor their bid/offer prices to existing market prices. Harry and we believe that an asynchronous-time market simulator such as JLMSim, which is capable of modeling the agents and market mechanisms behind observed prices, is much better than continuous-time models at representing the reality of markets.

So far, we have described JLMSim running in its dynamic analysis (DA) mode to simulate market behavior. More details about JLMSim running in the DA mode are given in Jacobs et al. (2004). JLMSim can also operate, in what we call capital markets equilibrium (CME) mode, to seek equilibrium expected returns, as we describe below.

Black and Litterman (1992) suggested a procedure to find equilibrium expected security returns that are consistent with a given covariance matrix and a specified market portfolio. The Black–Litterman (BL) procedure operates under the CAPM assumptions that investors can borrow all they want at the risk-free rate and that portfolios are constrained only by budget. It uses “reverse optimization” to compute equilibrium expected returns from the given covariance matrix and the specified market portfolio.

The BL procedure for estimating expected returns has as inputs: a covariance matrix; percentages of the market portfolio invested in various securities; views about expected returns for some, all, or none of the securities; and a parameter that serves to anchor the general level of expected returns. If the user supplies no views, the BL procedure produces Capital Market Equilibrium expected return estimates. These estimates are expected returns that would clear the market if every investor could borrow without limit at the risk-free rate.

Under the BL assumptions, investors are essentially unconstrained and can borrow all they want at the risk-free rate. Under these assumptions, the Tobin (1958) separation theorem applies and all investor portfolios lie on the straight capital market line (CML). Portfolios on the CML consist of various combinations of the riskless security and the *same* portfolio of risky securities.

In reality, contrary to the assumptions of the BL procedure, investors are constrained and cannot borrow all they want. Thus, investor portfolios do not all lie on the CML. Instead, they lie on the curved efficient frontier at positions determined by investor risk tolerances, and the compositions of the portfolios of risky securities differ from investor to investor. In such cases, the market portfolio may not even be efficient (see Markowitz 2005).

In the CME mode, JLMSim seeks capital market equilibrium expected returns for markets in which the CAPM assumptions do not necessarily hold. It allows users to solve for expected returns for markets in which investors cannot borrow, or have restricted borrowing, and in which investors can or cannot short. In other words, it can be used to seek equilibrium expected returns for any of the large variety of markets that can be simulated by JLMSim. Naturally, not all such markets are consistent with equilibrium solutions. Also, we have not explored the convergence properties of JLMSim for all such markets.

In CME mode, the objective of a run is to find capital market equilibrium expected security returns. These are found by adjusting securities' expected returns, thereby causing investors to change their portfolios in such a way that the aggregate of all investors' portfolios converges to given (or target) market portfolio weights. Generally speaking, if the weight of a security in the current market portfolio is above a given target weight, the simulator lowers the security's estimated expected return. If the current market weight is below the target weight, the simulator raises the security's estimated return.

More specifically, the user sets four nonnegative parameters $\{a_0, a_1, b_0, b_1\}$ used in the iterative adjustment of the equilibrium expected return estimates. For the i-th security, the iterative adjustment proceeds as follows. Let δ_i be the difference between the weight w_i^m of the security in the current market portfolio (computed from the aggregate of investors' holdings) and the weight w_i^t of the security in the target portfolio; i.e.,

$$\delta_i = w_i^m - w_i^t.$$

If the market weight is close enough to the target weight, so that

$$|\delta_i| < a_0 + a_1 w_i^\tau,$$

no action is taken. Otherwise, if δ_i is positive (negative), JLMSim subtracts (adds) $b_0 + b_1 w_i^t$ from (to) the current return estimate for security i. That is:

$$r_i \leftarrow r_i - \operatorname{sgn}(\delta_i)(b_0 + b_1 w_i^t).$$

The parameters a_0 and a_1 thus define a tolerance band around the target weights, outside of which adjustment of the return estimates is deemed necessary, and the parameters b_0 and b_1 define the degree to which the estimates should be adjusted.[13]

To create a realistic representation of market participants' holdings when running JLMSim in CME mode, the user can provide several investor templates that would place representative portfolios on various parts of the efficient frontier, and not just on the CML. With such placement, the BL assumptions are no longer satisfied. Therefore, the BL procedure would not provide correct equilibrium expected returns.

In contrast, JLMSim does provide correct results under these circumstances. The estimated equilibrium expected returns at the end of a CME run are the returns that are consistent with the given market portfolio and given covariance matrix. Furthermore, they are consistent with realistic assumptions regarding limits on investors' ability to borrow. Specific examples are provided in Jacobs et al. (2010).

[13] Some subtleties regarding the homogeneity of JLMSim's return estimates are discussed in Jacobs et al. (2010).

Those interested in finding out more about JLMSim, or experimenting with it, can download it from http://www.jacobslevy.com/jlmsim.

In conclusion, we hope that over time and with input from the finance community, JLMSim will develop into a simulator with which researchers can create realistic dynamic models of the market. Potentially, these models could help to test the effects on securities' prices of real-world events such as changes in investment strategy or regulatory policy. Other examples may include examining the effects on markets of various levels of passive portfolio management, or of leverage. Still others may include investigating the impact of institutional structures (such as minimum tick sizes, or the use of crossing networks), or regulatory policies (including, e.g., capital gains taxation and circuit breakers). Further, while it is still under development, JLMSim can already be used to compute capital market equilibrium returns under fairly realistic constraints that would make the same problem analytically intractable.

19.5 Conclusion: The Theory, Practice, and Future of Investing

In the 1950s, Harry's seminal ideas on portfolio theory, rigorously analytical and math-intensive, were met with skepticism by many in an investment industry that concentrated for the most part on stock picking. Subsequent critics claimed Harry's theories could not be translated into actual portfolios. Others touted ostensibly superior approaches such as semivariance analysis or portfolio insurance.

Yet Harry's mean–variance analysis remains – used and useful. One reason for the theory's longevity is its adaptability to practice. As Harry notes in his foreword to Jacobs and Levy (2000), "mean-variance analysis should not be considered a black box that can be set on automatic and allowed to run portfolios on its own." Human judgment is critical, and pure theory must be shaped by real-world considerations. In his ideas, as in his life, Harry bridges the gap between theory and practice.

From our most recent work with Harry on market simulation, we know that he remains as intrigued today by the workings of financial markets as he must have been in his graduate student days of the 1940s. His curiosity then led to the creation of modern portfolio theory. Now his collaboration with us looks forward to the creation of a practical model of the workings of the market. Not a small part of the workings of the market today reflect Harry's ideas on mean–variance analysis.

References

F. Black, and R. Litterman. Global Portfolio Optimization. *Financial Analysts Journal*, 48(5):28–43, September/October 1992.

M. Bloch, J. Guerard, H.M. Markowitz, G.P. Todd, and G. Xu. A Comparison of Some Aspects of the U.S. and Japanese Equity Markets. *Japan and the World Economy*, 5(1):3–26, May 1993.

K.J. Cohen, and J.A. Pogue. An Empirical Evaluation of Alternative Portfolio-Selection Models. *Journal of Business*, 40(2):166–193, April 1967.

E.J. Elton, M.J. Gruber, and M.W. Padberg. Simple Criteria for Optimal Portfolio Selection. *Journal of Finance*, 31(5):1341–1357, December 1976.

E.J. Elton, M.J. Gruber, and M.W. Padberg. Simple Criteria for Optimal Portfolio Selection: Tracing Out the Efficient Frontier. *Journal of Finance*, 33(1):296–302, March 1978.

B.I. Jacobs. *Capital Ideas and Market Realities: Option Replication, Investor Behavior, and Stock Market Crashes*. Blackwell, Malden, MA, 1999.

B.I. Jacobs. Risk Avoidance and Market Fragility. *Financial Analysts Journal*, 60(1), January/February 2004.

B.I. Jacobs. Tumbling Tower of Babel: Subprime Securitization and the Credit Crisis. *Financial Analysts Journal*, 65(1), January/February 2009.

B.I. Jacobs, and K.N. Levy. Disentangling Equity Return Regularities: New Insights and Investment Opportunities. *Financial Analysts Journal*, 44(3):18–43, 1988a.

B.I. Jacobs, and K.N. Levy. Calendar Anomalies: Abnormal Returns at Calendar Turning Points. *Financial Analysts Journal*, 44(6), 1988b.

B.I. Jacobs, and K.N. Levy, Forecasting the Size Effect. *Financial Analysts Journal*, 45(3), 1989.

B.I. Jacobs, and K.N. Levy. *Equity Management: Quantitative Analysis for Stock Selection*. McGraw-Hill, New York, NY, 2000.

B.I. Jacobs, and K.N. Levy. Enhanced Active Equity Portfolios are Trim Equitized Long-Short Portfolios. *Journal of Portfolio Management*, 33(4):19–25, Summer 2007.

B.I. Jacobs, K.N. Levy, and H.M. Markowitz. Financial Market Simulation. *Journal of Portfolio Management*, 30th Anniversary Issue:142–152, September 2004.

B.I. Jacobs, K.N. Levy, and H.M. Markowitz. Portfolio Optimization with Factors, Scenarios, and Realistic Short Positions. *Operations Research*, 53(4):586–599, July/August 2005.

B.I. Jacobs, K.N. Levy, and H.M. Markowitz. Trimability and Fast Optimization of Long-Short Portfolios. *Financial Analysts Journal*, 62(2):36–46, March/April 2006.

B.I. Jacobs, K.N. Levy, and H.M. Markowitz. Simulating Security Markets in Dynamic and Equilibrium Modes. *Forthcoming Financial Analysts Journal*, 2010.

B.I. Jacobs, K.N. Levy, H.M. Markowitz, and D. Starer. Equivalence and Relative Computational Complexity of the Critical Line and Elton-Gruber-Padberg Algorithms. Technical Report, Jacobs Levy Equity Management, 2007.

B.I. Jacobs, K.N. Levy, and D. Starer. On the Optimality of Long-Short Strategies. *Financial Analysts Journal*, 54(2):40–51, March/April 1998.

B.I. Jacobs, K.N. Levy, and D. Starer. Long-Short Portfolio Management: An Integrated Approach. *Journal of Portfolio Management*, 25(2):23–32, Winter 1999.

H. Levy, and H.M. Markowitz. Approximating Expected Utility by a Function of Mean and Variance. *American Economic Review*, 69(3):308–317, June 1979.

H.M. Markowitz. Portfolio Selection. *Journal of Finance*, 7(1):77–91, 1952.

H.M. Markowitz. The Optimization of a Quadratic Function Subject to Linear Constraints. *Naval Research Logistics Quarterly*, III:111–133, 1956.

H.M. Markowitz. *Portfolio Selection: Efficient Diversification of Investments*. Basil Blackwell, Cambridge, MA, 1959.

H.M. Markowitz. *Mean-Variance Analysis in Portfolio Choice and Capital Markets*. Basil Blackwell, Cambridge, MA, 1987.

H.M. Markowitz. Market Efficiency: A Theoretical Distinction and So What? *Financial Analysts Journal*, 61(5):17–30, September/October 2005.

W.F. Sharpe. A Simplified Model for Portfolio Analysis. *Management Science*, 9(2):277–293, January 1963.

J. Tobin. Liquidity Preference as Behavior Towards Risk. *The Review of Economic Studies*, 25(2):65–86, February 1958.

J.L. Treynor, and F. Black. How to Use Security Analysis to Improve Portfolio Selection. *Journal of Business*, 46(1):66–86, January 1973.

Chapter 20
Evaluating Hedge Fund Performance: A Stochastic Dominance Approach

Sheng Li and Oliver Linton

20.1 Introduction

Over the last decade, the number of hedge funds has risen by about 20% per year to reach around 11,000 in 2007. The amount of assets under management of the hedge fund industry has increased from around $40 billion in 1990 to an estimated $2,200 billion in 2007. Since hedge funds typically use leverage, the positions that they take in the financial markets are large enough to move markets around the world. The rapid growth in hedge funds reflects the increasing importance of this alternative investment category for institutional investors and wealthy individual investors.

Correspondingly, identifying hedge fund managers with superior skills and refining the traditional portfolio management tools to optimize investments in a large universe of hedge funds have also become challenging tasks in portfolio management. If the top hedge fund performance can be explained by superior skills owned by managers and not by luck, we would expect that top performance of such managers persists. However, there is little consensus on hedge fund performance persistence in the empirical finance literature. A number of studies find that hedge fund performance only persist at short term (1–3 months) which might be due to hedge funds' illiquid exposure and there is no evidence of performance persistence at annual horizons [see Getmansky et al. (2004), Brown et al. (1999), Agarwal and Naik (2000), Liang (2000), Bares et al. (2003), Boyson and Cooper (2004), and Baquero et al. (2005)]. On the contrary, more recent study by Kosowski et al. (2006) finds that sorting hedge funds on Bayesian alphas yields a 5.5% per year increase inthe alpha of the spread between the top and bottom hedge fund deciles. Hedge

S. Li (✉)
Department of Finance, London School of Economics, Houghton Street, London WC2A 2AE, UK
e-mail: s.li4@lse.ac.uk

O. Linton
Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, UK

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_20,
© Springer Science+Business Media, LLC 2010

fund performance persists at annual horizon. Using a novel GMM procedure to estimate alpha for hedge fund managers, Jagannathan et al. (2006) find evidence of hedge fund managers' performance persistence over 3-year horizons.

More practical issue facing hedge fund investors is how to construct an efficient hedge fund portfolio or add hedge funds to the existing portfolio. The standard mean–variance approach to portfolio allocation, which is founded on the assumption of normal distributions and an objective function of maximizing risk-adjusted return, is inadequate when dealing with portfolios of hedge funds. A number of studies [see Lo (2001) and Amin and Kat (2003)] have shown that risk characteristics of hedge funds are substantially different from those of traditional investment pools because hedge fund managers usually employ highly dynamic trading strategy and use short selling, leverage, concentrated investments, and derivatives. Specifically, hedge fund returns are not normally distributed and exhibit significant skewness and kurtosis. They also tend to display significant coskewness with the returns on other hedge funds as well as equity. Mean–variance models ignore these higher moments of the return distribution, and thus fail to take into consideration the benefits of funds that occasionally surprise on the upside while they also underestimate the risk of funds that have asymmetric downside risk. Despite the weakness of mean–variance frame work, it still dominates in practical hedge fund portfolio management. The Sharpe ratio is commonly used to quantify the risk-return trade-off. Amenc et al. (2004) report that only 2% of the European multimanagers pay attention to skewness and kurtosis; while 84% of multimanager funds consider that volatility is of major concern to their clients and 82% consider Sharpe ratio as an important indicator. A number of studies also address the issue of including hedge funds in standard institutional portfolios in mean–variance portfolio optimization [see Amenc and Martellini (2002), Brunel (2004), Kat (2005), and Till (2005)].

Another strand of literature develops different frameworks for hedge fund allocation, which incorporate a variety of investment objectives, particularly investor preferences for skewness and kurtosis of returns, into portfolio optimization models. Using a Polynomial Goal Programming (PGP) optimization model, Davies et al. (2005) solve for multiple competing hedge fund allocation objectives within a mean–variance–skewness–kurtosis framework and analyze different impacts of various hedge fund strategies on the distribution of optimal portfolio. Morton et al. (2006) study hedge fund allocation issue by assuming a family of utility functions which are a weighted sum of the probability of achieving a benchmark and expected regret relative to another benchmark. They then use a Monte Carlo method to obtain a solution to the related portfolio optimization model. Alexander and Dimitriu (2004) develop a portfolio construction model by selecting funds according to their ranking of alpha estimated with factor models. They then allocate selected funds using constrained minimum variance optimization.

In this paper, we introduce a more general and flexible framework for hedge fund asset allocation – stochastic dominance (SD) theory. Our approach utilizes statistical tests for stochastic dominance to compare the returns of hedge funds. The theory of stochastic dominance [see Hadar and Russell (1969), Hanoch and Levy (1969), Rothschild and Stiglitz (1970), and Whitmore (1970)] provides a systematic framework for analyzing economic behavior under uncertainty. We form hedge fund portfolios by using SD criteria. We then examine the out-of-sample performance of these hedge fund portfolios. Compared to both randomly selected hedge fund portfolio and mean–variance efficient hedge fund portfolio, our results show that fund selection method based on SD criteria greatly improves the performance of hedge fund portfolio.

Our framework relying on stochastic dominance has several advantages. First, we are able to use the information embedded in the entire empirical return distributions of hedge funds instead of a finite set of sample statistics. Second, while mean–variance analysis is consistent with the expected utility theory only under relatively restrictive assumptions about investor preferences or the statistical distribution of the investment returns, SD criteria do not require a full parametric specification of investor preferences, but rather rely on general preference assumptions which are intuitively close to the real objectives of investors, for example, nonsatiation in the case of first-order stochastic dominance (FSD) and risk aversion in the case of second-order stochastic dominance (SSD). This is important because the view of investors towards various hedge funds depends crucially on their investment objectives and risk preferences.

The reminder of the paper is organized as follows. Section 20.2 introduces stochastic dominance framework. Section 20.3 describes the data and reports the results of empirical analysis and a comparison of performance of various hedge fund portfolios constructed by using different criteria. Section 20.4 gives the concluding remarks.

20.2 Stochastic Dominance

Stochastic dominance theory provides a possible comparison relationship between two stochastic distributions. Stochastic dominance relations offer a general decision rule for decision making when facing the choice between random payoffs, given that the utility functions share some common characteristics such as nonsatiation or risk aversion. In this paper, we test for the first- and second-orders of stochastic dominance.

Let X_1 and X_2 be two outcome variables. Let $\mathcal{U}_1$ denote the class of all von Neumann–Morgenstern type utility functions, u, such that $u' \geq 0$, (increasing). Also, let $\mathcal{U}_2$ denote the class of all utility functions in $\mathcal{U}_1$ for which $u'' \leq 0$ (strict concavity). Let $F_1(x)$ and $F_2(x)$ denote the cumulative distribution functions, respectively.

Definition 20.1. X_1 First-Order Stochastic Dominates X_2, denoted $X_1 \succeq_{FSD} X_2$, if and only if:

(1) $E[u(X_1)] \geq E[u(X_2)]$ for all $u \in \mathcal{U}_1$, with strict inequality for some u; Or
(2) $F_1(x) \leq F_2(x)$ for all x with strict inequality for some x.

Definition 20.2. X_1 Second-Order Stochastic Dominates X_2, denoted $X_1 \succeq_{SSD} X_2$, if and only if either:

(1) $E[u(X_1)] \geq E[u(X_2)]$ for all $u \in \mathcal{U}_2$, with strict inequality for some u; Or:
(2) $\int_{-\infty}^{x} F_1(t)dt \leq \int_{-\infty}^{x} F_2(t)dt$ for all x with strict inequality for some x.

For any two outcomes i, j define

$$\delta_{ij} = \sup_{x \in \mathcal{X}} F_i(x) - F_j(x),$$

where $\mathcal{X}$ is contained in the supports of X_i, X_j. Fund i dominates fund j if $\delta_{ij} \leq 0$. If a fund X_1 second order dominates fund X_2 then no risk averse individual would prefer X_2 to X_1. First-order dominance of one outcome by another is even stronger: If a fund X_1 first order dominates fund X_2 then no individual who prefers more wealth to less would prefer X_2 to X_1. First-order dominance implies second-order dominance. Note that these concepts do not require the existence of moments of the underlying outcomes unlike mean–variance analysis. Furthermore, both relations are transitive, i.e., if X_1 dominates X_2 and X_2 dominates X_3 then X_1 dominates X_3. However, neither relation denotes a full ordering, only a partial ordering. That is, we may not be able to rank two outcomes at all according to either relation. In such cases, either one can say that one is indifferent between the two investments or one can impose more preference structure to discriminate between them. One possibility is to increase the dominance order to third order or fourth order etc., which reduces the set of noncomparability. Alternatively one can then supplement the partition induced by the dominance relation by some additional criterion like Sharpe ratio. In practice, although FSD implies strong relationship between two outcomes, it is not very discerning because the cumulative distributions of net returns of the two investment alternatives often intersect, in which case FSD cannot discriminate between the alternatives. For decision making under risk, more important is SSD. If investors are risk averse and prefer more to less, SSD could be used to choose between two outcomes.

In empirical analysis, stochastic dominance analysis requires the comparison of the probability distributions of two outcomes which are unknown and must be estimated from available data. Various statistical tests for the existence of SD orders have been developed. Several tests proposed earlier [for example Anderson (1996) and Davidson and Duclos (2000)] compare the distribution functions only at a fixed number of arbitrarily chosen points. In general, comparisons using only a small number of arbitrarily chosen points will have low power if there is a violation of the inequality in the null hypothesis on some subinterval lying between the evaluation points used in the test. More recent tests proposed by Barrett and Donald (2003) and Linton et al. (2003) compare the two distributions at all points in the sample.

20.3 Empirical Results

20.3.1 Description of the Data

In this section, we provide an empirical analysis of hedge fund database under Stochastic Dominance framework. The database used in this paper covers the period January 1994 to December 2007 and was provided by the Center for International Securities and Derivatives Markets (CISDM). It has two parts: a total of 1,269 live hedge funds and 1,760 dead hedge funds. To reduce survivorship bias, we include both live and dead funds in our analysis. Each set consists of a performance file, containing monthly net-of-fee returns, total net assets, and net asset values, and a fund information file, containing fund name, strategy type, management fees, and other supplementary details. We select only those funds with at least 2 years of monthly observations. We analyze seven fund strategies, namely Merger Arbitrage, Distressed Securities, Equity Hedge, Market Neutral, Convertible Arbitrage, Fixed Income Arbitrage, and Global Macro. Table 20.1 lists summary statistics of the hedge funds from CISDM database during the January 1994 to December 2007 period. For each strategy, the table lists the number of funds and means and standard deviations of basic summary statistics.

20.3.2 Results

To compare hedge fund returns using stochastic dominance concepts, our procedure includes two steps. First, we take into account the systematic risk exposure of hedge funds and obtain the risk-adjusted returns of hedge funds. Then we test for FSD and SSD relations among risk-adjusted hedge fund returns relying on Linton et al. (2003) statistical test.

Table 20.1 Summary statistics of hedge fund returns

	Mean		SD		Skewness		Kurtosis		$\hat{\rho}_1$%	
Category	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Market neutral	0.90	0.71	2.98	2.19	0.07	1.29	5.79	7.09	6.49	17.66
Equity hedge	1.18	0.69	4.55	2.73	0.19	0.86	5.49	3.72	12.48	13.51
Distressed securities	0.99	0.60	3.62	3.31	0.12	1.25	6.92	4.49	20.65	19.67
Merger arbitrage	0.91	0.72	2.97	3.59	−0.27	1.15	6.37	5.18	14.56	17.42
Convertible arbitrage	0.92	0.64	2.39	3.49	−0.21	1.34	6.58	5.32	32.02	19.81
Fixed income arbitrage	0.61	0.46	2.31	1.64	−1.02	1.98	9.99	11.57	15.17	20.14
Global macro	0.81	1.00	4.85	3.64	0.18	0.99	5.07	3.99	4.98	16.37

This table presents means and standard deviations of basic summary statistics for funds in the CIDSM database over the sample period January 1994 to December 2007. SD denotes standard deviations. $\hat{\rho}_1$% and $\hat{\rho}_2$% denote first-order and second-order autocorrelation, respectively

Risk adjustments for hedge fund returns are difficult due to their use of derivatives and dynamic trading strategies. Commonly used methods include using hedge fund indices and factor models. More recently, a number of studies [see Kat and Palaro (2005)] argue that sophisticated dynamic trading strategies involving liquid futures contracts can replicate many of the statistical properties of hedge fund returns. Hasanhodzic and Lo (2006) estimate linear factor models for individual hedge funds using six common factors and find that for certain hedge fund style categories, a significant fraction of funds' expected return can be captured by common factors.

Here we use as performance benchmarks the seven-factor model developed by Fung and Hsieh (2004). The Fung and Hsieh (2004) factors are S&P 500 return minus risk-free rate (SNPMRF), Wilshire small cap minus large cap return (SCMLC), change in the constant maturity yield of the 10-year Treasury (BD10RET), change in the spread of Moody's Baa minus the 10-year Treasury (BAAMTSY), bond PTFS (PTFSBD), currency PTFS (PTFSFX), and commodities PTFS (PTFSCOM), where PTFS denotes primitive trend following strategy. Fung and Hsieh (2004) show that their factor model strongly explains variation in individual hedge fund returns.

In order to obtain risk-adjusted performance of hedge funds, we regress the net-of-fee monthly excess return (in excess of the risk-free rate) of a hedge fund on the seven-factor model.

$$R_{i,t} = \alpha_i + \beta_i^\top Z_t + \epsilon_{i,t}, \tag{20.1}$$

where β_i represents the risk exposure of fund i at month t to the various factors and Z_t is the monthly value of different factors. The risk-adjusted return of fund i at month t is calculated as:

$$\hat{a}_{i,t} = R_{i,t} - \hat{\beta}_i^\top Z_t = \hat{\alpha}_i + \hat{\epsilon}_{i,t}, \tag{20.2}$$

where $R_{i,t}$ is the net-of-fee monthly excess return of fund i in month t, $\hat{\beta}_i$ is the estimated risk exposure for fund i, and Z_t is the value of the various factors at month t. We compute the risk-adjusted returns $\hat{a}_{i,t}$ as the sum of the intercept $\hat{\alpha}_i$ and the residual $\hat{\epsilon}_{i,t}$ of (20.1). We plot the distribution of the intercept α in Fig. 20.1.

We next conduct an analysis of the distributions of risk-adjusted returns of the funds, with a view to establishing stochastic dominance orderings. For each fund i we compute the empirical c.d.f. and integrated c.d.f. (denoted s.d.f.) as follows

$$\widehat{F}_i(x) = \frac{1}{T_i}\sum_{t=1}^{T_i} 1(X_{it} \le x),$$

$$\widehat{S}_i(x) = \int_{-\infty}^{x} \widehat{F}_i(x')\mathrm{d}x' = \frac{1}{T_i}\sum_{t=1}^{T_i}(x - X_{it})1(X_{it} \le x),$$

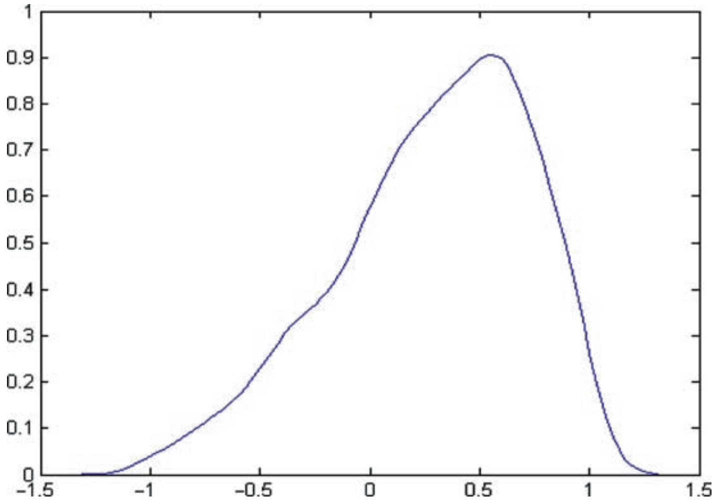


Fig. 20.1 Cross-sectional estimates of alpha of hedge funds in our sample

where $X_{it} = \hat{a}_{i,t}$ is risk-adjusted return. We say that a fund i is first order dominated if for some fund j

$$\max_{1\leq \ell \leq L} \widehat{F}_j(x_\ell) - \widehat{F}_i(x_\ell) < 0,$$

where $x_1, \ldots, x_L$ is a grid of points contained in the union of the supports of the distributions. Likewise, for second-order dominance.

Let $\mathcal{F}_D = \{i : i \text{ is first order dominated}\}$ and let $\mathcal{F}_U$ be the complement of this set in the full set of funds, likewise define $\mathcal{S}_D$ and $\mathcal{S}_U$. Clearly, $\mathcal{F}_D \subseteq \mathcal{S}_D$ and so $\mathcal{S}_U \subseteq \mathcal{F}_U$.

We compute the set of all funds that are undominated across all pairwise comparisons. We then construct a portfolio of all undominated funds. To examine the out-of-sample performance of undominated funds, we construct portfolios of funds $\mathcal{S}_U$ on January 1 each year (from 1999 to 2007), based on stochastic dominance orders of risk-adjusted hedge fund returns estimated over the prior 5 years. The portfolios are equally weighted monthly, so the weights are readjusted whenever a fund disappears.[1] We also construct the portfolio of first-order dominated funds for comparison purpose. Given the economic intuition of stochastic dominance that any risk-averse individual should choose funds in $\mathcal{S}_U$ and any investor who prefer more to less should not choose funds in $\mathcal{F}_D$, we expect portfolio of funds in $\mathcal{S}_U$ exhibit much better performance than portfolio of funds in $\mathcal{F}_D$.

To compare stochastic dominance tests with mean–variance tests, we also apply mean and variance efficient criteria to risk-adjusted returns of hedge funds. We con-

[1] Under a pessimistic scenario, the money invested into disappeared hedge funds cannot be recovered. Hence we assume -100% return to a fund during the month after it disappears from the database and zero returns thereafter.

struct portfolios of mean–variance efficient funds on January 1 each year (from 1999 to 2007), based on means and variances of risk-adjusted returns of funds estimated over the prior 5 years. A fund is defined as a mean–variance efficient fund if no other funds have both higher means and lower variances than this fund. Hence, funds are selected by comparing only two summary statistics: the mean and the variance, which represents the distribution of risk-adjusted hedge fund returns.

A number of studies find that hedge fund portfolio return properties vary substantially with the number of hedge funds included in the portfolio. See Amin and Kat (2002), Davies et al. (2003) and Alexander and Dimitriu (2004). A hedge fund portfolio including only ten funds will typically have significantly higher variance than a similar hedge fund portfolio containing 100 funds. Therefore, to assess the robustness of SD analysis, we also construct 20 representative portfolios containing the same number of funds as in $\mathcal{S}_U$ for each year (from 1999 to 2007). The funds in portfolios are randomly selected. We then compute the average statistics of 20 portfolios.

Table 20.2 reports the number of hedge funds held by portfolios for each year and Table 20.3 reports summary statistics and alphas of portfolios constructed using different criteria. Alpha is estimated using the seven-factor model. As we can see in Table 20.2, the number of funds in $\mathcal{S}_U$ is around 30 which is close to those of mean–variance efficient funds while the number of funds in $\mathcal{F}_D$ is substantially larger, ranging from 417 to 1,009. French et al. (2005) examine the current fund of hedge funds universe and find that funds of hedge funds report holding between 1 and 200 underlying funds generally hold 10–30 with close to 20 on average. Hence, the number of holdings in $\mathcal{S}_U$ and mean–variance efficient sets is actually close to practitioner standards. Amin and Kat (2003) also find that the optimal size of well-diversified hedge fund portfolios is in the range of 15–20.

Table 20.2 Number of funds

Year	Total N of funds	N of funds in F_D	N of funds in S_U	N of MV efficient funds	N of funds in both S_U and MV efficient
1999	591	417	16	24	13
2000	713	438	22	24	16
2001	810	505	15	18	13
2002	940	513	43	37	30
2003	1,048	535	38	28	21
2004	1,183	804	27	35	25
2005	1,338	1,009	18	29	18
2006	1,297	993	10	15	10
2007	1,267	686	22	22	19

This table reports the numbers of funds held in portfolios of second-order undominated funds (S_U), mean–variance efficient funds, first-order dominated funds (F_D) over the sample period: January 1999 to August 2007. The portfolios are constructed on January 1 each year. N is the number of funds

Table 20.3 Summary statistics of returns for representative portfolios

	Mean	Max	Min	Std. dev.	Skew	Kurtosis	Alpha (pct/year)	t-Stat of alpha
Funds in S_U	1.08	6.96	−5.63	2.53	0.71	6.42	14.28	3.88
MV efficient funds	0.95	6.74	−8.09	2.34	−0.47	4.92	12.36	4.40
Funds in F_D	0.11	5.96	−5.89	2.20	−0.32	3.34	1.32	0.51
Funds randomly picked	0.58	8.33	−10.97	2.92	−0.91	6.54	4.32	2.12
CISDM fund of funds index	0.63	4.24	−2.10	1.1	0.03	4.06	6.36	6.90

This table reports summary statistics of returns for portfolios of second-order undominated funds (S_U), mean–variance efficient funds, first-order dominated funds (F_D), randomly selected funds over the sample period: January 1999 to December 2007. The portfolios are constructed on January 1 each year

According to Table 20.3, the mean return of portfolio of funds in $\mathcal{S}_U$ is 1.08 which is substantially larger than those of other portfolios. The first two moments of returns provide a great deal of the information about the investment outcome set of portfolios, but not everything. We find the skewness for portfolio of funds in $\mathcal{S}_U$ is 0.71 which is larger than for other portfolios. Positive skewness means essentially that the big outcomes are on the upside so there is relatively little chance of large negatives. From a variety of points of view positive skewness is desirable.

Moreover, the portfolio of funds in $\mathcal{S}_U$ generates an alpha of 14.4% per year. The t-statistics shows this alpha is statistically significant. The alphas of portfolios of funds in $\mathcal{F}_D$ and mean–variance efficient funds are also statistically significant but much lower than the alpha of portfolio of funds in $\mathcal{S}_U$. The alpha of the randomly picked funds portfolio is the lowest.

Figure 20.2 plots the time series of returns of these representative portfolios. We also plot the cumulative returns[2] of representative portfolios and CISDM fund of funds index in Fig. 20.3. As we can see from the figure, the portfolio constructed by using SD criterion achieves a much higher cumulative return than those of other portfolios.

To further investigate the nature of the stochastic dominance approach, we establish stochastic orders within each style category. We then repeat the above performance analysis. Table 20.4 reports the results for each category. We find that overall the portfolio of funds in $\mathcal{S}_U$ display superior performance. In particular, the portfolio of funds in $\mathcal{S}_U$ in the Merger arbitrage category achieve relative higher alpha than that of mean–variance efficient funds. For Global macro funds, the performance of stochastic dominance approach is worse than that of mean–variance approach.

[2] The cumulative return is the compound return of the series: $CumR_n = \left(\prod_{i=1}^{n}(1+r_i)\right) - 1$, where $r_1, \ldots, r_n$ is a monthly return series.

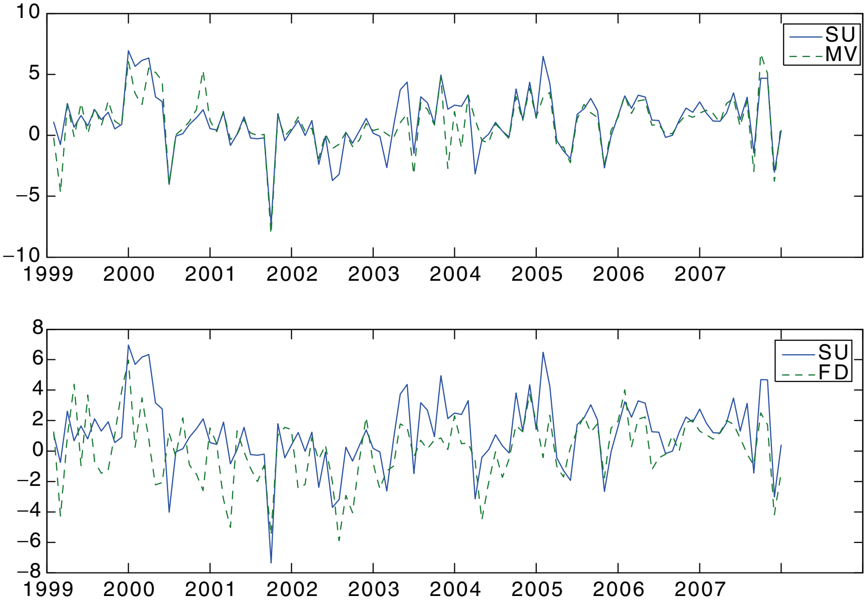


Fig. 20.2 Returns of portfolios of second undominated funds (SU), mean–variance efficient funds (MV) and first-order dominated funds (FD). The portfolios are constructed on January 1 each year from 1999 to 2007

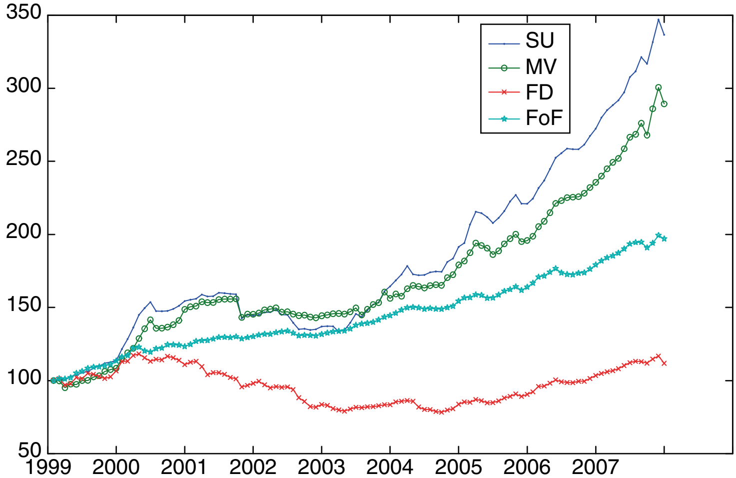


Fig. 20.3 Cumulative returns of portfolios of second undominated funds (SU), mean–variance efficient funds (MV), first-order dominated funds (FD). The portfolios are constructed on January 1 each year from 1999 to 2007. For comparison purpose, we also plot cumulative returns of CISDM fund of fund index (01/1999 to 12/2007)

Table 20.4 Summary statistics of returns for representative portfolios within styles

	Mean	Max	Min	Std. dev.	Skew	Kurtosis	Alpha (pct/year)	t-stat of alpha
Equity hedge								
Funds in S_U	1.25	9.28	−8.15	2.93	0.42	5.61	14.40	3.09
MV efficient funds	0.92	6.62	−8.74	2.84	−0.80	4.57	10.20	2.46
Funds randomly picked	0.83	10.84	−9.88	3.19	−0.29	5.13	9.36	2.71
Equity neutral								
Funds in S_U	0.68	5.23	−4.43	1.38	0.34	5.35	8.26	3.89
MV efficient funds	0.61	6.31	−3.62	1.47	−0.02	5.77	7.32	3.38
Funds randomly picked	0.60	7.18	−9.42	2.09	−0.81	10.25	6.95	2.79
Merger arbitrage								
Funds in S_U	0.63	12.61	−8.01	2.18	1.19	14.06	7.56	2.88
MV efficient funds	0.47	8.52	−6.44	1.89	−0.31	8.42	4.96	1.98
Funds randomly picked	0.38	4.01	−11.56	1.88	−2.86	17.66	4.25	1.96
Distressed securities								
Funds in S_U	0.84	4.68	−5.04	1.52	−0.69	5.52	9.69	4.96
MV efficient funds	0.83	4.25	−5.23	1.42	−0.67	5.33	9.60	5.38
Funds randomly picked	0.69	8.80	−12.78	2.63	−1.48	17.74	8.45	2.39
Convertible arbitrage								
Funds in S_U	0.90	5.20	−7.28	1.47	−0.84	12.27	10.08	5.97
MV efficient funds	0.88	4.32	−4.93	1.13	−1.03	9.67	9.98	7.32
Funds randomly picked	0.15	7.19	−8.75	2.74	−2.23	19.72	2.11	0.89
Fixed income arbitrage								
Funds in S_U	0.77	3.06	−2.39	0.92	0.29	4.34	9.36	6.73
MV efficient funds	0.64	2.50	−3.38	0.89	−1.25	7.57	7.80	5.95
Funds randomly picked	0.52	4.96	−9.93	1.99	−2.20	14.56	5.26	2.48
Global macro								
Funds in S_U	0.45	9.69	−7.67	2.50	−0.11	5.56	4.32	1.49
MV efficient funds	0.49	7.77	−6.05	2.39	−0.07	3.97	6.12	2.08
Funds randomly picked	0.39	18.46	−13.51	3.69	0.76	19.75	4.20	0.99

For each style category, this table reports summary statistics of returns for portfolios of second-order undominated funds (S_U), mean–variance efficient funds, randomly selected funds over the sample period: January 1999 to December 2007. The portfolios are constructed on January 1 each year

20.4 Concluding Remarks

In this paper, we introduce a general and flexible framework for hedge fund performance evaluation and asset allocation. Our approach utilizes recent advances in statistical tests for stochastic dominance. The approach is able to recognize and use the information embedded in the nonnormal return distributions of hedge funds. To illustrate the method's ability to work with nonnormal distributions, we form hedge fund portfolios by using SD criteria and examine the out-of-sample performance of these hedge fund portfolios. Compared to performance of portfolios of randomly selected hedge funds and mean–variance efficient hedge funds, our results show that fund selection method based on SD criteria greatly improves the performance of hedge fund portfolio. The mean return of portfolio of funds in S_U is substantially larger than those of other portfolios. We also find that the skewness for portfolio of funds in S_U is 0.71 which is larger than for other portfolios. Positive skewness is desirable because it means essentially that the big outcomes are on the upside so there is relatively little chance of large negatives. Mean–variance optimization models do not necessary achieve this result. Different specifications of investor preferences will result in considerable differences in the impact of skewness on optimal hedge fund allocations.[3]

There are a number of potential areas for improvement. First, the equal weighting of undominated funds can be replaced by more targeted weighting based on some univariate performance criterion like Sharpe ratio. Second, we could look at higher order dominance or asymmetric dominance notions like Prospect or Markowitz dominance. Third, we could take account of sampling variation in constructing the set of undominated funds by including those funds that are within some distance [controlled according to a statistical criterion like Type 1 error using the results of Linton et al. (2003)] from being dominated. This would enlarge the set of undominated funds and it may not improve performance out of sample. Finally, although our SD test shows ability in distinguishing good funds from bad funds, it is restricted to pairwise comparison of a finite number of choice alternatives and it has limitations with full diversification possibilities. The problem is that the ordering of the outcomes of a diversified portfolio of funds cannot be determined in a straightforward way from the orderings of the individual funds. Therefore, the ordering of each portfolio has to be determined individually. A number of recent studies recently developed Linear Programming (LP) tests for SD that do fully account for diversification.[4] Post and van Vliet (2006) develop tests for SSD and TSD efficiency that are embedded in the Generalized Method of Moments (GMM) framework. This test has superior statistical properties to the above LP tests and is a serious rival to the dominant mean–variance tests. We leave the application of these tests to hedge funds as future research.

References

Agarwal, V. and N. Y. Naik, 2000, Multi-Period Performance Persistance Analysis of Hedge Funds, *Journal of Financial and Quantitative Analysis*, 35, 327–342.

Alexander, C. and A. Dimitriu, 2004, The Art of Investing in Hedge Funds: Fund Selection and Optimal Allocations, Working paper, ISMA Centre, University of Reading.

Amenc, N. and L. Martellini, 2002, Portfolio Optimization and Hedge Fund Style Allocation Decisions, *The Journal of Alternative Investments*, 5, 7–20.

[3] See Brockett and Kahane (1992) and Cremers et al. (2005).

[4] See Post (2003), Kuosmanen (2004), Post and Levy (2005), and Post and van Vliet (2006).

Amenc, N., J. Giraud, L. Martellini and M. Vaissie, 2004, Taking a Close Look at the European Fund of Hedge Funds Industry: Comparing and Contrasting Industry Practices and Academic Recommendation, *Journal of Alternative Investments*, 7, 59–69.

Amin, G.S. and H. Kat, 2003, Hedge Fund Performance 1990–2000: Do the Money Machines Really Add Value?, *Journal of Financial and Quantitative Analysis*, 38, 251–274.

Anderson, G.J., 1996, Nonparametric Tests of Stochastic Dominance in Income Distributions, *Econometrica,* 64, 1183–1193.

Baquero, G., J. Ter Horst and M. Verbeek, 2005, Survival, Look-Ahead Bias and the Persistence in Hedge Fund Performance, *Journal of Financial and Quantitative Analysis*, 40, 493–517.

Bares, P.-A., R. Gibson and S. Gyger, 2003, Performance in the Hedge Fund Industry: An Analysis of Short and Long-Term Persistence, *Journal of Alternative Investments*, 6, 25–41.

Barrett, G. and S. Donald, 2003, Consistent Tests for Stochastic Dominance, Unpublished manuscript, University of New South Wales.

Boyson, N. M. and M. J. Cooper, 2004, Do Hedge Funds Exhibit Performance Persistence? A New Approach, Working paper.

Brockett, P. L. and Y. Kahane, 1992, Risk, return, skewness and preference, *Management Science*, 38, 851–866.

Brown, S. J., W. N. Goetzmann and R. G. Ibbotson, 1999, Offshore Hedge Funds: Survival and Performance 1989–1995, *Journal of Business*, 72, 91–118.

Brunel, J., 2004, Revisiting the Role of Hedge Funds in Diversified Portfolios, *The Journal of Wealth Management*, 7, 35–48.

Cremers, J., M. Kritzman and S. Page, 2005, Optimal Hedge Fund Allocation, *Journal of Alternative Investments*, 31, 70–81.

Davidson, R. and J-Y. Duclos, 2000, Statistical inference for stochastic dominance and for the measurement of poverty and inequality, *Econometrica*, 68, 1435–1464.

Davies, R. J., H. M. Kat and S. Lu, 2005, Fund of Hedge Funds Portfolio Selection: A Multiple-Objective Approach, Cass Business School Research Paper.

French, C. W., D. B. Ko and D. N. Abuaf, 2005, Diversification and Persistence in Hedge Funds. Available at SSRN: http://ssrn.com/abstract=850304

Fung, W. and D. Hsieh, 2004, Hedge Fund Benchmarks: A Risk-Based Approach, *Financial Analysts Journal*, 60, 65–80.

Getmansky, M., A. W. Lo and I. Makarov, 2004, An Econometric Model of Serial Correlation and Illiquidity in Hedge Fund Returns, *Journal of Financial Economics*, 74, 529–609.

Hadar, J. and W. R. Russell ,1969, Rules for Ordering Uncertain Prospects, *American Economic Review*, 59, 25–34.

Hanoch, G. and H. Levy, 1969, The Efficiency Analysis of Choices Involving Risk, *Review of Economic Studies*, 36, 335–346.

Hasanhodzic, J. and A. W. Lo, 2006, Can Hedge-Fund Returns Be Replicated?: The Linear Case, Working paper.

Jagannathan, R., A. Malakhov and D. Novikov, 2006, Do Hot Hands Persist Among Hedge Fund Managers? An Empirical Evaluation, NBER Working paper no. W12015.

Kat, H., 2005, Integrating Hedge Funds into the Traditional Portfolio, *The Journal of Wealth Management*, 7, 51–57.

Kat, H. and H. Palaro, 2005, Who Needs Hedge Funds? A Copula-Based Approach to Hedge Fund Return Replication, Alternative Investment Research Centre Working paper No. 27, Cass Business School, City University, London.

Kosowski, R., N. Y. Naik and M. Teo, 2006, Do Hedge Funds Deliver Alpha? A Bayesian and Bootstrap Analysis, *Journal of Financial Economics*, 84, 229–264.

Liang, B., 2000, Hedge Funds: The Living and the Dead, *Journal of Financial and Quantitative Analysis*, 35, 309–326.

Linton, O., E. Maasoumi and J. Whang ,2003, Consistent Testing for Stochastic Dominance Under General Sampling Schemes, *Review of Economic Studies*, 72, 735-765.

Lo, A. W., 2001, Risk Management for Hedge Funds: Introduction and Overview, *Financial Analysts Journal*, 57, 16–33.

Morton, D., E. Popova and I. Popova, 2006, Efficient fund of hedge funds construction under downside risk measures, *Journal of Banking and Finance*, 30, 503–518.

Okunev, J. and D. White, 2003, Hedge Fund Risk Factors and Value at Risk of Credit Trading Strategies, Working paper, Principal Global Investors and University of New South Wales.

Popova, I., D. P. Morton and E. Popova, 2003, Optimal hedge fund allocation with aysmmetric preferences and distributions, Technical Report, The University of Texas at Austin.

Popova, I., D. P. Morton and E. Popova, 2006, Efficient Fund of Hedge Fund Construction Under Downside Risk Measures, *Journal of Banking and Finance*, 30, 503–518.

Post, T., 2003, Stochastic Dominance in Case of Portfolio Diversification: Linear Programming Tests, *Journal of Finance*, 58, 1905–1931.

Post, T. and H. Levy, 2005, Does Risk Seeking Drive Stock Prices? A Stochastic Dominance Analysis of Aggregate Investor Preferences and Beliefs, *Review of Financial Studies*, 18, 925–953.

Post, T. and P. van Vliet, 2006, Downside Risk and Asset Pricing, *Journal of Banking and Finance*, 30, 823–849.

Post, G. T. and P. J. P. M. Versijp, 2007, Multivariate test for stochastic dominance efficiency of a given portfolio, *Journal of Financial and Quantitative Analysis*, (ISSN 0022-1090), Vol. 42(2), 489–515.

Rothschild, M. and J. E. Stiglitz, 1970, Increasing Risk I: A Definition, *Journal of Economic Theory*, 2, 225–243.

Till, H., 2005, On the Role of Hedge Funds in Institutional Portfolios, *The Journal of Alternative Investments*, 8, 77–89.

Whitmore, G. A., 1970, Third-Degree Stochastic Dominance, *American Economic Review*, 60, 457–459.

Chapter 21
Multiportfolio Optimization: A Natural Next Step

Martin W.P. Savelsbergh, Robert A. Stubbs, and Dieter Vandenbussche

21.1 Introduction

Mean–variance optimization of a single portfolio, as introduced by Markowitz (1952, 1959), is well studied and well understood. Its influence can be found in many branches of the quantitative finance community (Fabozzi et al. 2002). Advances in mathematical programming techniques have not only allowed for the fast and reliable solution of large-scale mean–variance optimization problems, but have also allowed for the incorporation of many relevant business considerations, such as limits on the number of names, threshold position levels, and even long/short ratios. As a result, several commercial software vendors now provide asset management solutions based on mean–variance optimization.

The success of these solutions has naturally led to additional demands from the market place and the exploration of extensions to the mean–variance optimization framework by academics as well as software vendors. A good example is the thriving business of separately managed accounts. Separately managed accounts were introduced by asset management firms to accommodate clients who wanted to meet specific objectives, which did not fit within the constrictions of a mutual fund investment. The freedom of choice of professional managers, portfolio customization, objective investment advice for a set fee, diversification (or concentration if the client so chooses), tax efficiency, and general flexibility have made separate accounts popular among informed investors. To be able to effectively and efficiently cope with many similar, but not identical portfolios, managers (often rebalancing these portfolios more often too) need process automation and decision support. However, as portfolio managers realized and experienced quickly, it is not simply a matter of automating the workflow and employing optimization tools to rebalance multiple portfolios. Typically, separately managed accounts are customized versions of a particular investment style or model. As a result, when these

M.W.P. Savelsbergh (✉)
Axioma, Inc. & Georgia Institute of Technology, Atlanta, USA
e-mail: martin.savelsbergh@gatech.edu

R.A. Stubbs and D.Vandenbussche
Axioma, Inc., Atlanta, USA

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_21,
© Springer Science+Business Media, LLC 2010

portfolios are rebalanced, a significant overlap exists in the assets selected to be traded. As it is common to pool the trades of different accounts for common execution, the price that is used during the rebalancing process of the accounts tends to be an underestimate (sometimes a substantial underestimate) of the price at execution time due to the market impact of the *combined* volume of the trades. (Market impact cost is typically modeled as a nonlinear strictly convex function of the trade size.) To be able to more accurately model the market impact cost of trades, the portfolios should be rebalanced simultaneously in one large portfolio optimization problem, that is, a multiportfolio optimization problem.

Hence, the development and use of multiportfolio optimization is a natural next step in the evolution of mean–variance optimization. Of course, the value of multi–portfolio optimization is not restricted to more accurately accounting for the market impact cost of pooled trades, as it may also lead to improved returns by properly capturing the interaction of the portfolios. For example, if the trade size of an asset is limited based on the average daily trade volume (ADV), then if a portion of the "allocation" to one portfolio goes unused, another account may benefit as it can trade more.

Our primary goal in this chapter is to demonstrate that using multiportfolio optimization is a practically viable option and to provide insight into the benefits that can be derived from doing so. For that purpose, we perform a computational study using realistic data and realistic portfolio sizes.

Our secondary goal is to highlight potential issues concerning fairness that may arise when simultaneously optimizing multiple portfolios, as it is the fiduciary responsibility of the portfolio manager to ensure that the best interest of each account is considered and that no account benefits to the detriment of another.

Two concepts, from the microeconomic theory of oligopolies, have been used to analyze market equilibriums: *collusive* solutions and *Cournot-Nash equilibrium* solutions. In a collusive solution, the total welfare is maximized, that is, we maximize the sum of the objective functions of the individual accounts. In a Cournot-Nash equilibrium solution, rather than colluding to maximize total welfare, each account optimizes its own objective assuming the trade decisions of all other accounts that participate in the pooled trading have been made and are known and fixed. We introduce a multiportfolio rebalancing framework that encompasses both settings, that is, it allows simultaneous rebalancing of multiple accounts while ensuring that each account is optimized according to either the collusive solution or the Cournot-Nash equilibrium solution, and adheres to all account-specific constraints as well as any additional constraints that span across multiple accounts. Furthermore, we examine, by means of a few small examples, the advantages and disadvantages of both types of solutions.

The remainder of this chapter is organized as follows. In Sect. 21.2, we elaborate on multiportfolio optimization, highlight the computational challenges, and present the results of an extensive computational study. In Sect. 21.3, we study and analyze fairness issues related to multiportfolio optimization. Finally, in Sect. 21.4, we discuss other settings in which multiportfolio optimization may be used effectively and comment on future research directions and opportunities.

21.2 Multiportfolio Optimization

Multiportfolio optimization provides a platform for scaling modern portfolio rebalancing techniques from the individual account level, at which they are originally specified, to the level of a pooled, simultaneous rebalancing. Modern portfolio rebalancing techniques that rely on optimization use trading costs as an important component of the objective function. When large positions in an equity asset are bought or sold, the price of the security can be affected as the trades are being executed. Therefore, the average price at which the trade is executed is often worse than expected. This implicit cost of trade execution is known as the *market impact cost*. Market impact cost is typically modeled using a strictly convex nonlinear function of the trade volume, such as a quadratic function or to a power between 1.5 and 2 (see Almgren et al. 2005). As a result, the market impact cost computed when each account is rebalanced independently *underestimates* the actual market impact cost that will be incurred when trading the whole pool of accounts together. That is, the actual market impact cost of trading multiple accounts is typically much greater than the sum of the estimated market impact cost of trading each account separately. Therefore, any approach that ignores the aggregate market impact cost may result in a reduction in realized returns for some or all of the accounts that are being optimized. Considering the aggregate effect of market impact cost of the combined trades and properly accounting for them during the rebalancing process ensures that the resulting portfolios are optimal under the conditions in which they are actually managed and traded. Because some accounts may be affected by trading cost underestimation more than others, ignoring the aggregate market impact may also result in unintended biases, or *unfair trading*. Multiportfolio optimization, however, does not eliminate the possibility of unfair trading. We will investigate unintended biases resulting from multiportfolio optimization in Sect. 21.3.

To date, the aggregate effect of trading several accounts in a pool is often ignored when the accounts are optimized independently. Even when the aggregate effects of joint trading are taken into account, they are considered in an ad-hoc, or heuristic at best, process. One technique frequently used is to aggregate similar accounts, rebalance them together as a single representative account, and then allocate trades to the individual accounts on a pro-rata basis. Alternatively, one could approximate the true market impact cost by assuming that all accounts to be rebalanced will execute similar trades and hence adjusting the incremental trade costs appropriately. Another common procedure applied in practice is to spread the rebalancing of accounts over the rebalancing cycle in a round-robin fashion to reduce the market impact of one account on another. This approach, of course, may also produce suboptimal solutions, increase account dispersion, and potentially bias certain accounts. An account that is rebalanced later in the cycle may suffer from stale expected returns or trade assets that have already been affected by the permanent market impact of the other accounts that have been rebalanced and traded earlier in the cycle. Spreading the rebalancing of accounts throughout the rebalancing cycle may also generate dispersion of returns. A more sophisticated approach is to "optimize" the multiportfolio optimization problem using an iterative process in which each account is optimized independently, trades are accumulated across all accounts, and constraints and/or objective functions for each individual account are then adjusted to better represent the true cost of aggregate trading. Of course, if one is not careful, the process may take a large number of iterations to converge, if it converges at all. Furthermore, the solution is almost surely suboptimal and may bias a particular type of account.

True multiportfolio optimization provides an elegant and exact approach for determining an optimal rebalancing for multiple accounts at once. For this, a comprehensive model is set up, capturing the individual requirements for each account on separate sets of variables, and introducing linking constraints over the variables of the individual accounts to get a more realistic estimate of the market impact costs. These linking constraints, one per asset, compute the value of the sum of all trades for the asset and use this value in the objective function to compute the market impact cost for the totally traded amount.

For a more detailed explanation, consider a set of accounts $\mathcal{A}$. Let h_i represent the vector of initial holdings for account $i \in \mathcal{A}$ and w_i represent the vector of optimized holdings for account $i \in \mathcal{A}$. Then $t_i = w_i - h_i$ represents the vector of amounts traded for each asset in account $i \in \mathcal{A}$. The objective function of each individual account $i \in \mathcal{A}$ is $f_i(w_i) + t_i^{\mathrm{T}} c_i(t_i)$, where $f_i(w_i)$ represents the expected return and $c_i(t_i)$ is a vector function giving the market impact cost per unit of currency traded for each asset. The function used to predict the market impact cost, $t_i^{\mathrm{T}} c(t_i)$, is commonly used and is nonlinear and strictly convex in the amounts traded. The vector function $c(\cdot)$ is often independent for each asset and expressed as a polynomial of the form t_k^p for each asset k where p is a rational number between 0.5 and 1 (see Almgren et al. 2005). The constraints for each individual account $i \in \mathcal{A}$ are represented as a vector of inequalities $g_i(w_i) \geq 0$. The optimization problem for each individual account can then be written as follows:

$$
\begin{aligned}
\text{maximize } & f_i(w_i) - t_i^{\mathrm{T}} c(t_i) \\
\text{subject to } & t_i = w_i - h_i \\
& g_i(w_i) \geq 0.
\end{aligned}
$$

When the accounts are optimized simultaneously, the trade amounts of each individual account are combined to form the pooled trade amount, and the problem becomes:

$$
\begin{aligned}
\text{maximize } & \textstyle\sum_{i \in \mathcal{A}} f_i(w_i) - t^{\mathrm{T}} c(t) \\
\text{subject to } & t = \textstyle\sum_{i \in \mathcal{A}} (w_i - h_i) \\
& g_i(w_i) \geq 0 \ \forall i \in \mathcal{A}.
\end{aligned}
$$

Thus, the market impact cost is based on the total net trade value. Multiportfolio optimization, therefore, identifies a rebalancing for all accounts simultaneously while taking into account realistic costs for the pooled trades. It should be noted that the trade costs are *not* allocated to individual accounts. Finding a fair allocation of the commonly incurred costs of pooled trades to the separate client accounts may be challenging.

To study the benefits of using multiportfolio optimization for rebalancing of separately managed accounts, we have conducted two computational experiments, which we describe in the next sections.

21.2.1 Computation Study: Market Impact Objective

In the first experiment, we compare the expected returns of a set of accounts when each of the accounts is rebalanced separately to the expected returns when the accounts are rebalanced simultaneously, where expected returns are the returns minus the trading costs. More specifically, we consider 10 accounts with similar, but non-identical, percent holdings. The universe of assets as well as the benchmark is the S&P 600 and each account has about 250 holdings. The accounts are managed using the following strategy. The objective is to maximize expected return less market impact cost. Furthermore, shorting is not allowed, a threshold holding limit of 0.1% is enforced, an active risk limit of 3% is imposed on the portfolio, active holding limits of $\pm 3\%$ are imposed on the individual assets, and active industry limits are set to $\pm 5\%$. To assess the market impact cost of trades, we use the Goldman Sachs Shortfall Model, which generates pretrade estimates of the likely impact of a proposed trade on the price of an asset. When the accounts are rebalanced individually, the total *anticipated* expected return (over all accounts) is \$237,295,868, whereas the total *actual* expected return (over all accounts) is only \$227,679,160 due to the underestimation of the market impact cost. When the accounts are rebalanced simultaneously, the total expected return is \$229,784,310. This represents an increase in expected return of about 1%. In Table 21.1, we show in more detail how this increase in expected returns is achieved. We present for each account, the size of the account, the initial expected return, the anticipated and actual objective function value as well as the anticipated and actual market impact cost when the account is rebalanced individually, and the objective value and market impact cost when the accounts are rebalanced simultaneously (where expected returns are presented as a percentage of the portfolio value). An examination of the results shows that the accounts do not benefit equally from simultaneous rebalancing. In fact, the objective value for Account 1 & 2 decreases (from 4.76% to 4.72%), whereas the objective value would have increased (from 2.17% to 4.76% and from 2.19% to 4.76%, respectively) if the accounts were rebalanced individually. The smaller accounts (in terms of account value) are "subsidizing" the larger accounts. This demonstrates that multiportfolio optimization has to be applied carefully and that methods may need to be developed to ensure that the way the gains are divided over the accounts is perceived as fair, because otherwise undesirable situations may arise.

Table 21.1 Multiportfolio optimization

			Individually optimized				Simultaneously optimized	
Account	Value	Initial exp. return (%)	Anticipated obj. value (%)	Anticipated MI-cost (%)	Actual obj. value (%)	Actual MI-cost (%)	Obj. value (%)	MI-cost (%)
A1	100MM	2.17	5.01	0.16	4.76	0.44	4.72	0.31
A2	200MM	2.19	5.00	0.20	4.76	0.44	4.72	0.31
A3	300MM	1.52	4.93	0.26	4.67	0.52	4.71	0.32
A4	400MM	1.45	4.88	0.30	4.67	0.51	4.72	0.31
A5	500MM	1.18	4.85	0.33	4.85	0.54	4.68	0.35
A6	600MM	1.19	4.83	0.35	4.64	0.54	4.68	0.35
A7	700MM	0.84	4.77	0.40	4.57	0.60	4.62	0.41
A8	800MM	0.85	4.75	0.42	4.57	0.60	4.62	0.41
A9	900MM	0.86	4.73	0.43	4.56	0.60	4.61	0.42
A10	1000MM	0.93	4.73	0.43	4.56	0.60	4.61	0.42

21.2.2 Computational Study: ADV Constraints

In the second experiment, we eliminate the market impact objective and introduce trade limits based on the average daily trade volumes (ADV). These constraints are applied to the aggregate trades across the accounts, which implies, like the market impact objective, that the trades of one account can impact the trades of another. We consider the same set of accounts and manage the accounts using essentially the same strategy as before. The objective is simply to maximize expected return. Furthermore, when the accounts are optimized simultaneously, trade size limits of 20% of ADV are imposed for all assets in the universe, and when the accounts are optimized individually each account is "allocated" a fraction of the trade size limit based on relative size of the account value, that is, if account i has value V_i, then the fraction of the trade size limit allocated to account i is $\frac{V_i}{\sum_i V_i}$. When the accounts are rebalanced individually, the total expected return (over all accounts) is \$94,259,555. When the accounts are rebalanced simultaneously, the total expected return is \$107,030,375. This represents an increase in expected return of more than 13%. In Table 21.2, we show in more detail how this increase in expected returns is achieved. We present for each account, the size of the account, the initial expected return, the expected return when the account is rebalanced individually, and the expected return when the accounts are rebalanced simultaneously (where expected returns are presented as a percentage of the portfolio value). A number of observations can be made. First, the expected returns after simultaneous rebalancing are almost the same for all accounts (i.e., 2.17 %). More importantly, however, the accounts do not benefit equally from simultaneous rebalancing. In fact, the expected return for Account 2 decreases (from 2.19% to 2.18%), whereas the expected return would have increased (from 2.19% to 2.99%) if the account was rebalanced individually. Again, the smaller accounts (in terms of account value) are "subsidizing" the larger accounts. As this situation is undesirable, we have also conducted a simultaneous rebalancing in which we impose for each account that the expected

Table 21.2 Multiportfolio optimization with ADV limits

		Expected return		
Account	Value	Initial (%)	Individually optimized (%)	Simultaneously optimized (%)
A1	100MM	2.17	2.97	2.18
A2	200MM	2.19	2.99	2.18
A3	300MM	1.52	2.36	2.18
A4	400MM	1.45	2.29	2.17
A5	500MM	1.18	1.97	2.17
A6	600MM	1.19	1.98	2.17
A7	700MM	0.84	1.65	2.16
A8	800MM	0.85	1.69	2.16
A9	900MM	0.86	1.68	2.16
A10	1000MM	0.93	1.70	2.16

Table 21.3 Multiportfolio optimization with ADV limits and expected return bounds

		Expected return		
Account	Value	Initial (%)	Individually optimized (%)	Simultaneously optimized (%)
A1	100MM	2.17	2.97	2.97
A2	200MM	2.19	2.99	2.99
A3	300MM	1.52	2.36	2.36
A4	400MM	1.45	2.29	2.29
A5	500MM	1.18	1.97	2.08
A6	600MM	1.19	1.98	2.07
A7	700MM	0.84	1.65	2.08
A8	800MM	0.85	1.69	2.08
A9	900MM	0.86	1.68	2.07
A10	1000MM	0.93	1.70	2.07

return has to be at least the expected return that can be achieved when the account is rebalanced individually. In this case, the total expected return, when the accounts are rebalanced simultaneously is $106,826,757.77, which still represents an increase of more than 13%. In Table 21.3, we show, once more, how this increase in expected returns is achieved. As we can see, it is still the case that the smaller accounts are "subsidizing" the increase of expected returns of the larger accounts, but at least it is not done to the detriment of the smaller accounts.

Our computational study demonstrates the potential benefits of multiportfolio optimization for separately managed accounts, but it also shows the challenges associated with sharing the gains. In the next section, we investigate the issue of fairness in a more abstract setting.

21.3 Fairness

Although striving for optimality is the primary consideration when solving multiportfolio optimization problems, it may also be desirable, or even necessary, to obtain unbiased solutions, that is, solutions in which the trades are allocated fairly across accounts.

To the best of our knowledge, O'Cinneide et al. (2006) were first to analyze the fairness or bias of the simultaneous rebalancing of multiple accounts using multiportfolio optimization. They consider a multiportfolio optimization problem where the objective function is the sum of the objective functions of each individual account. In microeconomic theory of oligopolies, this solution is referred to as the *collusive* solution [see Varian (1984) and Henderson and Quandt (1980)]. The collusive solution maximizes total welfare over all the accounts and O'Cinneide et al. (2006) assert that it is fair to each individual account, because the solution obtained is the same as the one that would have been obtained if each account was competing in an open market for liquidity. We show below that in certain situations, however, it may happen that certain accounts may be better off acting alone rather than by participating in the collusive solution. When participating in a collusion, some of the accounts may make sacrifices for the good of others to maximize total welfare.

An alternative to identifying the set of portfolios that maximizes the total welfare is identifying the set of portfolios that form a Cournot-Nash equilibrium. Like the collusive solution, the Cournot-Nash equilibrium solution originates in the study of the behavior of oligopolies in the microeconomic literature. However, rather than colluding to maximize total welfare, each participant optimizes its own objective assuming the trade decisions of all accounts that participate in the pooled trading have been made and are fixed. Such a solution has the property that no account would have an incentive to unilaterally change its trades. A Cournot-Nash equilibrium solution, however, may not be Pareto optimal, which means that there exists a set of trades where each account performs at least as good as with the equilibrium solution and at least one account obtains improved expected performance. We show below how the Cournot-Nash equilibrium solution can be obtained from the solution to a single multiportfolio optimization problem.

In the examples used in this section, we will assume, without loss of generality, that each account has an all-cash initial position and is not allowed to short assets. Again let $\mathcal{A}$ be the set of accounts to be rebalanced and let w_i represent the vector of asset holdings for account i ($i \in \mathcal{A}$) in units of currency. As a result, the cost of executing the trades for account i ($i \in \mathcal{A}$) is given by $w_i^{\mathrm{T}} c(w_1; \ldots; w_{|\mathcal{A}|})$, where $c(w_1; \ldots; w_{|\mathcal{A}|})$ is the market impact cost function. To simplify the argument, we further assume, without loss of generality, that $c(w_1; \ldots; w_{|\mathcal{A}|})$ is linear and represented as $\Omega \sum_{j \in \mathcal{A}} w_j$, where Ω is a symmetric positive semidefinite matrix.

Now that we have all of the necessary information, let us consider the different alternatives available for solving the multiportfolio optimization problem. We will examine the optimality of each and also discuss any biases found in each of the proposed approaches.

Optimizing Accounts Independently. First, consider a very simple example where the objective function for account i is to maximize its utility (represented by the expected return less portfolio variance) less market impact cost without any constraints. Then, the optimization problem used to determine the portfolio for account i can be written as follows:

$$\underset{w_i \geq 0}{\text{maximize}} \ \alpha^{\mathrm{T}} w_i - \rho w_i^{\mathrm{T}} Q w_i - w_i^{\mathrm{T}} \Omega w_i, \tag{21.1}$$

where α is a vector of expected returns, Q is a covariance matrix of asset returns, and ρ is a risk-aversion parameter. The optimal solution of (1) can be written as

$$\begin{aligned} w_i &= (2\rho Q + 2\Omega)^{-1} (\alpha + \lambda_i), \\ \lambda_{ik} w_{ik} &= 0 \text{ for each asset } k, \\ \lambda_i &\geq 0. \end{aligned} \tag{21.2}$$

In this setting, each account is acting independently and is unaware of the effect on the market impact cost that the other accounts have. If each account is optimized according to the same strategy, then the actual market impact cost to account i is $w_i^{\mathrm{T}}\Omega\left(\sum_{j\in\mathcal{A}} w_j\right)$ rather than the $w_i^{\mathrm{T}}\Omega w_i$ that was estimated. The actual market impact cost to each account is $100(|\mathcal{A}|-1)\%$ greater than expected, for example, if the number of accounts is 100, then the actual market impact cost is 9900% greater than expected.

In the scenario above, information of the trades being executed is not shared amongst the accounts being rebalanced. It is clear that each account is under-optimized because of the under-estimation of market impact cost. However, this in itself does not mean that any particular account is treated unfairly. The question is whether or not any particular account has a consistent advantage over others in such a scenario.

Suppose that there are only two accounts, one with \$100M and one with \$10B in total assets, and that both start from cash positions. Furthermore, suppose that our problem is to maximize expected return less market impact cost subject to being fully invested and a 10% risk constraint. Each account is optimized completely independently of the other. The results of the portfolio rebalancings are summarized in Table 21.4, where "Expected Market Impact" and "Expected Objective" represent the values computed for the rebalancing (using only the trade amounts of the single account) while "Actual Market Impact" and "Actual Objective" represent the true values when the market impact cost is based on the combined trade amounts of both accounts. Note the large decrease in the objective value for the \$100M account. Because the trades of the large account have a much greater effect on the market impact cost of the smaller account than vice-versa, the smaller account is consistently being hurt. The smaller account may argue that it is treated "unfairly."

In summary, when accounts are optimized independently each account incurs a greater market impact cost than expected. Furthermore, unintended biases can occur and can result in certain accounts being consistently worse off.

Cournot-Nash Equilibrium Solution. The market impact cost associated with jointly trading multiple accounts can be modeled as a strictly convex nonlinear function of all trades across the accounts. As a result, it is not appropriate to treat

Table 21.4 Summary statistics of the independent account rebalancing example

Property	Account 1	Account 2
Size (\$)	100M	10B
Predicted risk (%)	10.0	10.0
Expected return (%)	14.54	11.72
Expected market impact (%)	0.1821	2.5192
Actual market impact (%)	4.4403	2.5618
Expected objective (%)	14.36	9.20
Actual objective (%)	10.10	9.16
Change in objective (%)	−29.7	−0.4

the market impact in a single-account optimization as if the account is the only account being traded. As we have seen, doing so underestimates the true trading cost of rebalancing the account. Instead, the market impact caused by all accounts being optimized simultaneously needs to be considered.

Next assume that the rebalancing problem for each account is “made aware” of the trades of the other accounts that are being pooled together for execution. Then, the information about these trades can be incorporated in the rebalancing problem for an account. Let us assume that the trades of all other accounts have been determined and do not change, that is, the values of $w_j \geq 0$ are known and fixed for each account $j \in \mathcal{A} \setminus i$. Under this scenario, the optimization problem for account i can be modeled as follows:

$$\underset{w_i \geq 0}{\text{maximize}} \;\; \alpha^{\mathrm{T}} w_i - \rho w_i^{\mathrm{T}} Q w_i - w_i^{\mathrm{T}} \Omega \left(\textstyle\sum_{j \in \mathcal{A}} w_j \right). \tag{21.3}$$

The optimal solution to (3) is given by

$$\begin{aligned} w_i &= (2\rho Q + 2\Omega)^{-1} \left(\alpha + u_i - \Omega \textstyle\sum_{j \in \mathcal{A} \setminus i} w_j \right), \\ u_{ik} w_{ik} &= 0 \text{ for each asset } k, \\ u_i &\geq 0. \end{aligned} \tag{21.4}$$

The solution satisfying these conditions for each account is the Cournot-Nash equilibrium solution. An attractive property of this approach is that the actual market impact cost for each account is exactly what the account expected.

Now let us revisit the two-account example considered earlier. The results comparing the Cournot-Nash equilibrium solution to the individual account optimizations are summarized in Table 21.5. Note that Account 1 has a greater actual objective in the Cournot-Nash equilibrium solution than it does in the individual account optimization. Account 2 has a greater actual objective as well, though the difference is so small that it does not show up in the number of significant digits represented in the table. Furthermore, notice that the aggregate actual objective across both accounts is greater for the equilibrium solution than for the individual account optimizations.

Table 21.5 Summary statistics of the Cournot-Nash equilibrium example

	Individual solutions		Equilibrium solution	
Property	Account 1	Account 2	Account 1	Account 2
Size ($)	100M	10B	100M	10B
Predicted risk (%)	10.0	10.0	10.0	10.0
Expected return (%)	14.54	11.72	14.39	11.69
Expected market impact (%)	0.1821	2.5192	4.1543	2.5329
Actual market impact (%)	4.4403	2.5618	4.1543	2.5329
Expected objective (%)	14.36	9.20	10.24	9.16
Actual objective (%)	10.10	9.16	10.24	9.16
Aggregate objective (%)	9.1651		9.1685	

The Cournot-Nash equilibrium solution approach is implicitly being sought by those quantitative portfolio managers that do consider the effects of other accounts on the market impact cost in their portfolio construction process. However, and because of the heuristic nature of most practical implementations, the true Cournot-Nash equilibrium is rarely found.

Throughout the above discussion, we have been using the *net alpha*, that is, the expected return minus the market impact cost, to compare the results for each multiaccount optimization approach. The Cournot-Nash equilibrium solution optimizes this net alpha objective for each account and dominates the solution obtained by optimizing each account individually, because the net alpha is optimized after trades for all accounts are considered.

In the collusive solution, to be discussed next, certain investors may find that it is possible to increase their net alpha by using the portfolio obtained by optimizing the account individually. Consequently, if we interpret the collusive solution as an equilibrium solution to some game, then the objective of each investor in this game is *not* the net alpha objective.

Collusive Solution. In the Cournot-Nash equilibrium solution, the optimization problem for each account assumes that the trades of the other accounts are known and fixed. This does not seem ideal because account i is considering the effect of the other accounts on its market impact cost, but not vice versa. To rectify this, we need to consider not only the market impact term, $w_i^{\mathrm{T}}\Omega\left(\sum_{j\in\mathcal{A}} w_j\right)$, but also the term $\left(\sum_{j\in\mathcal{A}\setminus i} w_j\right)^{\mathrm{T}}\Omega\left(\sum_{j\in\mathcal{A}} w_j\right)$. In the collusive approach, all investors collude to maximize the total welfare over all accounts. Therefore, the problem is a single joint optimization problem where the objective is to maximize the sum of the individual account objectives. This single optimization problem can be written as follows:

$$\underset{w_i\geq 0\ \forall i\in\mathcal{A}}{\text{maximize}}\ \sum_{j\in\mathcal{A}}\alpha^{\mathrm{T}}w_j-\rho\sum_{j\in\mathcal{A}}w_j^{\mathrm{T}}Qw_j-\left(\sum_{j\in\mathcal{A}}w_j\right)^{\mathrm{T}}\Omega\left(\sum_{j\in\mathcal{A}} w_j\right). \quad (21.5)$$

The optimal solution to (21.5) is given by

$$\begin{aligned} &w_i = (2\rho Q + 2\Omega)^{-1}\left(\alpha + v_i - 2\Omega\sum_{j\in\mathcal{A}\setminus i} w_j\right)\ \forall i \in \mathcal{A},\\ &v_{ik}w_{ik} = 0 \text{ for each asset } k \text{ and } \forall i \in \mathcal{A},\\ &v_i \geq 0\ \forall i \in \mathcal{A}. \end{aligned} \quad (21.6)$$

While the solution given by (21.6) maximizes the total utility, as represented by the sum of the individual utilities less market impact costs of all accounts, the question is whether or not this is fair to each individual account.

We consider the same two-account example and compare the collusive solution to the Cournot-Nash equilibrium solution. A summary of the results is given in Table 21.6. Note that Account 1 is better off using the Cournot-Nash equilibrium solution than the collusive solution; its expected net alpha is 107 basis

Table 21.6 Summary statistics of the collusive example

	Equilibrium solution		Collusive solution	
Property	Account 1	Account 2	Account 1	Account 2
Size ($)	100M	10B	100M	10B
Predicted risk (%)	10.0	10.0	10.0	10.0
Expected return (%)	14.39	11.69	11.69	11.69
Actual market impact (%)	4.1543	2.5329	2.5207	2.5207
Actual objective (%)	10.24	9.16	9.17	9.17
Aggregate objective (%)	9.1685		9.1732	

points more. The aggregate welfare across both accounts is greater in the collusive solution (although only minimally). So, the investors are colluding to increase the expected net alpha of Account 2 by approximately one basis point to the detriment of Account 1 for the sake of improving total welfare. In both the individual and Cournot-Nash equilibrium solutions, Account 1 is able to invest in higher alpha, but less-liquid, assets to improve its expected net alpha. Because Account 1 is better off in both of the other approaches, one might argue that Account 1 is being negatively biased in the collusive approach.

It is interesting to note that the collusive solution can also be obtained as a Cournot-Nash equilibrium solution by using an objective function that is different from the one used when rebalancing the account individually. Consider the following optimization problem:

$$\underset{w_i \geq 0}{\text{maximize}} \;\; \alpha^{\mathrm{T}} w_i - \rho w_i^{\mathrm{T}} Q w_i - w_i^{\mathrm{T}} \Omega w_i - 2 w_i^{\mathrm{T}} \Omega \left(\textstyle\sum_{j \in \mathcal{A} \setminus i} w_j \right). \qquad (21.7)$$

The optimal solution to (21.7) is given by

$$\begin{aligned} w_i &= (2\rho Q + 2\Omega)^{-1} \left(\alpha + u_i - 2\Omega \textstyle\sum_{j \in \mathcal{A} \setminus i} w_j \right), \\ u_{ik} w_{ik} &= 0 \text{ for each asset } k, \\ u_i &\geq 0. \end{aligned} \qquad (21.8)$$

When we write (21.8) for each account, we get the exact same optimality conditions as we did for the collusive solution given by (21.6). Note that each account has to believe that it must pay a market impact cost that results from its own trades as well as *twice* the trades of the other account. Because Account 2 is larger, the higher expected returns for the relatively illiquid assets that are available for Account 1 no longer appear attractive, because the rebalancing problem assumes that Account 2 is actually purchasing twice as much as it really is.

A Multiaccount Rebalancing Framework. We now describe a framework for multiaccount optimization that encompasses both the collusive and the Cournot-Nash equilibrium approaches.

Recall that $t_i = w_i - h_i$ represents the vector of amounts traded for each asset in account $i \in \mathcal{A}$ and that $t = \sum_{i \in A} t_i$. The objective function of each individual account i is $f_i(w_i) + t_i^{\mathrm{T}} c(t)$, where $f_i(w_i)$ represents the expected return (or a more complex utility function containing other terms, such as transaction costs, taxes, risk, etc.) and $c(\cdot)$ is a vector function giving the market impact cost per unit of currency traded. The constraints for each account i are written as a vector of inequalities $g_i(w_i) \geq 0$. Define T_i to be a vector that contains the net trade amount across all accounts except i. That is, $T_i = \sum_{j \in \mathcal{A} \setminus i} (w_j - h_j)$, where w_j and h_j are known for all $j \in \mathcal{A} \setminus i$.

Then the optimization problem for each account when the effect of trades of other accounts is considered can be written as follows:

$$\begin{aligned} \text{maximize } & f_i(w_i) - t_i^{\mathrm{T}} c(t) \\ \text{subject to } & t_i = w_i - h_i \\ & t = t_i + T_i \\ & g_i(w_i) \geq 0. \end{aligned} \tag{21.9}$$

To correctly optimize all accounts simultaneously, we solve an aggregate optimization problem involving all accounts. For the collusive approach, the aggregate optimization problem adds the objectives together while enforcing the constraints for each account. The collusive aggregate optimization problem can be written as follows:

$$\begin{aligned} \text{maximize } & \textstyle\sum_{i \in \mathcal{A}} f_i(w_i) - t^{\mathrm{T}} c(t) \\ \text{subject to } & t = \textstyle\sum_{i \in \mathcal{A}} (w_i - h_i) \\ & g_i(w_i) \geq 0 \ \forall i \in \mathcal{A}. \end{aligned} \tag{21.10}$$

The problem used to determine the Cournot-Nash equilibrium depends on the particular market impact function used. When the market impact function is quadratic, for example, and can be written as $\tilde{t}^{\mathrm{T}} \tilde{\Omega} \tilde{t}$, where

$$\tilde{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_{|\mathcal{A}|} \end{pmatrix} \text{ and } \tilde{\Omega} = \begin{bmatrix} \Omega & \frac{1}{2}\Omega & \cdots & \frac{1}{2}\Omega \\ \frac{1}{2}\Omega & \Omega & & \vdots \\ \vdots & & \ddots & \frac{1}{2}\Omega \\ \frac{1}{2}\Omega & \cdots & \frac{1}{2}\Omega & \Omega \end{bmatrix}, \tag{21.11}$$

then the Cournot-Nash equilibrium problem can be obtained by solving the following single multi-portfolio optimization problem:

$$\begin{aligned} \text{maximize } & \textstyle\sum_{i \in \mathcal{A}} f_i(w_i) - \tilde{t}^{\mathrm{T}} \tilde{\Omega} \tilde{t} \\ \text{subject to } & t_i = w_i - h_i \ \forall i \in \mathcal{A}, \\ & g_i(w_i) \geq 0 \ \forall i \in \mathcal{A}. \end{aligned} \tag{21.12}$$

Other market impact cost functions may require different solution approaches as it may not be possible to write them as a single objective function.

Problems (21.10) and (21.12) represent "generic" multiportfolio problems that are representative of practical instances, albeit simple enough to keep the mathematical arguments straightforward. In practice, more complex problems can be solved with the same approach. For example, we might allow for combinatorial constraints, and constraints on holding or trading across accounts. The formulations given in (21.9) and (21.10) assume that the market impact is modeled as a function of the net trade amount in each asset. Alternatively, market impact can be modeled as a function of the sum of the absolute values of buys and sells, the maximum of buys and sells, or other related functions. In each of these cases, a similar optimization problem can be created having the same desirable properties of optimality and computability. Combinatorial constraints such as thresholds or limits on the number of assets held can also be modeled, although some of the desirable properties of the optimal solution may be lost. Similarly, constraints across all accounts can be handled using the same approach, for example, a constraint on the total amount traded in each asset.

Problem (21.10), or the aforementioned more generic form, can be solved by the same methods used to solve single account problems. In our experience, such solution methods scale quadratically as a function of the number of accounts on a single processor. A decomposition-based approach could be used to enhance the performance of the optimizer, and achieve better than quadratic scaling, especially in the presence of a parallel architecture.

21.4 Final Remarks

We have shown that simultaneously optimizing multiple portfolios has to be done with care, because solutions may have unintended biases, that is, one portfolio may be favored over another. We have introduced and analyzed the collusive solution and the Cournot-Nash equilibrium solution. Both have their advantages and disadvantages and, as a result, it is impossible to argue that one has to be preferred over the other. Thus, careful analysis of the solution of multiportfolio optimization is a must.

On the positive side, we have also shown that realistic-size multiportfolio optimization problems can efficiently and effectively be solved with technology currently available in the market place. This technology should be of great value in the separately managed accounts business.

Another natural place to use multiportfolio optimization is in assisting investors that want to spread their investments over different management strategies (Bertsimas et al. 1999). The market offers a vast range of differently styled equity portfolios for investment. Each of these portfolios is typically managed according to a particular strategy. From an investor's perspective, this rich variety of investment possibilities provides an opportunity as well as a challenge. It offers the promise of increasing *overall* returns, but taking full advantage of the interactions between the different strategies is nontrivial. To accommodate such investors, firms may offer to manage a portfolio that is virtually subdivided into subportfolios, each with its own individual strategy. Such a division does not only enable flexible diversification across these strategies, but also allows for their interaction for mutual benefit. Investment controls can be specified on the subportfolio level as well as on the total portfolio level. For example, one can specify risk limits for the subportfolios as well as for the total portfolio or restrictions on the number of names held in the subportfolios as well as in the total portfolio. Independent simultaneous tracking of multiple benchmarks is also possible. Since the division is virtual, cross-trading between the subportfolios is allowed to satisfy controls and to achieve goals.

References

Robert Almgren, Chee Thum, Emmanuel Hauptmann, and Hong Li. Equity market impact. *Risk*, July 2005.

Dimitris Bertsimas, Christopher Darnell, and Robert Soucy. Portfolio construction through mixed-integer programming at Grantham, Mayo, Van Otterloo and Company. *Interfaces*, 29(1):49–66, 1999.

Colm O'Cinneide, Bernd Scherer, and Xiaodong Xu. Ensuring fairness when pooling trades in quantitative investment process. Technical report, Deutsche Asset Management, September 2006.

Frank J. Fabozzi, Francis Gupta, and Harry M. Markowitz. The legacy of modern portfolio theory. *Journal of Investing*, 11(3):7–22, 2002.

James M. Henderson and Richard E. Quandt. *Microeconomic Theory: A Mathematical Approach*. McGraw-Hill, Inc., third edition, 1980.

Harry M. Markowitz. Portfolio selection. *The Journal of Finance*, 7:77–91, 1952.

Harry M. Markowitz. *Portfolio Selection: Efficient Diversification of Investments*. Wiley, New Jersey, 1959.

Hal R. Varian. *Microeconomic Analysis*. W.W. Norton & Company, Inc., second edition, 1984.

Disclaimer: The multiportfolio optimization computational analysis described in this chapter was performed using Axioma Portfolio enhanced with the Goldman Sachs Shortfall Model ("Shortfall Model") for generating Shortfall estimates, and such Shortfall estimates may not be identical to Shortfall estimates generated with the Shortfall Model in a different operating environment. Such differences may arise due to processing requirements (e.g., portfolio optimization requirements) of the operating environment. Please contact Axioma for further information. Axioma is not an affiliate of Goldman, Sachs & Co.

The Shortfall Model produces information that is additional to, and compiled in a format differing from, that required under applicable regulations to which regulated financial institutions and broker-dealers are subject. Goldman, Sachs & Co. and/or its affiliates ("GS") have taken reasonable steps to ensure such information and/or formatting ("Materials"), as provided to Axioma, Inc. for use in this product is accurate and complete, but apprise the user that this information and/or formatting is being furnished "AS IS" as a courtesy. GS DISCLAIMS ALL REPRESENTATIONS AND WARRANTIES, EXPRESS OR IMPLIED, RELATING TO THE MATERIALS, INCLUDING BUT NOT LIMITED TO, WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, SECURITY AND

NON-INFRINGEMENT. The Materials are not an official report and, as such, GS is not responsible for any errors or omissions. GS does not represent that the Materials are accurate, complete and/or current, and does not accept liability for losses or damages arising from use of this information and/or formatting. GS SHALL UNDER NO CIRCUMSTANCES BE HELD LIABLE WHETHER IN CONTRACT, IN TORT, UNDER ANY WARRANTY OR ANY OTHER THEORY OF LIABILITY, FOR ANY INDIRECT, INCIDENTAL, SPECIAL, CONSEQUENTIAL OR PUNITIVE DAMAGES, INCLUDING BUT NOT LIMITED TO, LOSS OF REVENUE, LOSS OF PROFITS, LOSS OF OPPORTUNITY, BUSINESS DISRUPTION OR OTHER PECUNIARY LOSS ARISING OUT OF THE USE OF OR INABILITY TO USE THE MATERIALS, EVEN IF GS HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. In the event of any discrepancy between the Materials and the terms set forth in any official confirmation or statement with respect to a particular transaction or transactions, the terms of the official confirmation or statement shall govern.

Chapter 22
Alternative Model to Evaluate Selectivity and Timing Performance of Mutual Fund Managers: Theory and Evidence

Cheng-few Lee, Alice C. Lee, and Nathan Liu

22.1 Introduction

The investment on mutual funds has been extensively studied in finance. Over the last few decades, there has been a dramatic increase in the development of instruments measuring the performance of mutual funds. Early researchers (Treynor 1965; Sharpe 1966; and Jensen 1968) employed a one-parameter indicator to evaluate the portfolio performance. We can easily compare their performance by these estimated indicators. However, these studies assume the risk levels of the examined portfolios to be stationary through time.

Fama (1972) and Jensen (1972) pointed out that the portfolio managers may adjust their risk composition according to their anticipation of the market. Moreover, Fama (1972) suggested that the managers' forecasting skills can be divided into two parts: the selectivity ability and the market timing ability. The former is also known as micro-forecasting, involving identifying whether the stocks are over-valued relative to the general stocks or not. The latter is also called macro-forecasting, involving the forecast of future market return. In other words, the selectivity and market timing abilities of fund managers are viewed as important factors that decide the overall fund performance.[1]

Treynor and Mazuy (1966) used a quadratic term for the excess market return to test the market timing ability. It can be viewed as the extension of the Capital Asset Pricing model (CAPM). If the fund manager can forecast the market trend,

C.-f. Lee (✉)
Room 141, School of Business, Janice H. Levin Building, Rockafeller Road, Piscataway, NJ 08854-8054, USA
e-mail: lee@rbsmail.rutgers.edu.

A.C. Lee
State Street, Boston, MA, USA

N. Liu
Department of Finance, Feng Chia University, Taiwan, Republic of China.

[1] However, Brinson et al. (1991) found that selectivity and market timing abilities only have small influence on fund performance (<10%). The overall performance should be mostly decided by asset allocation between stock and bond markets.

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_22,
© Springer Science+Business Media, LLC 2010

he will change the proportion of the market portfolio in advance. Jensen (1972) developed the theoretical structure for the timing ability. Under the assumption of a joint normal distribution of the forecasted and realized returns, Jensen showed that the correlation between the managers' forecast and the realized return can be used to measure the timing ability. Bhattacharya and Pfleiderer (1983) extended Jensen's (1972) work and used a simple regression method to obtain accurate measures of selectivity and market timing ability. Lee and Rahman (1990) further corrected the estimation of parameters by a Generalized Least Squares (GLS) method. In addition, Henriksson and Merton (1981) used options theory, developed by Merton (1981), to explain the timing ability.

In this chapter, we empirically examined the mutual fund performance by using six models, proposed respectively by Treynor (1965), Sharpe (1966), Jensen (1968), Treynor and Mazuy (1966), Henriksson and Merton (1981), and Lee and Rahman (1990). There are 189 monthly returns ranging from January 1990 to September 2005 for a sample of 628 open-end equity funds used in the empirical study. In addition to examining the selectivity, timing, and overall performance, we also try to find a relationship between the estimated parameters and the real investment.

22.2 Methodologies

It is well known that return is not a satisfactory indicator for evaluating the performance; hence, it is necessary to consider the risk taken. Markowitz (1952) was the first to quantify the link that exists between return and risk, and also built the foundation of modern portfolio theory. Moreover, the Markowitz model contains the fundamental elements of the CAPM, which was also the basis for most of the models discussed in this chapter.

Treynor (1965) uses the concept of security market line,[2] drawn from the CAPM, to determine a coefficient β. Under the assumption of complete diversification of asset allocation, it means that we just have systematic risk measured by β. The Treynor index (TI), measuring the reward per unit of systematic risk for the portfolio, can be showed as follows:

$$TI = \frac{\overline{r}_p - r_f}{\beta_p}, \tag{22.1}$$

where $\overline{r}_p$ is the average return of the p^{th} mutual fund and r_f is defined as the risk-free rate. The numerator of Treynor index can be viewed as an excess return on the portfolio. This ratio is a risk-adjusted performance value. This indicator is suitable for valuing the performance of a well-diversified portfolio; this is because it just takes the systematic risk into account.

[2] At equilibrium, all assets are located on this line.

In contrast, Sharpe (1966) argues the phenomenon that the fund managers favor fewer stocks. Therefore, it is impossible to diversify the individual risks completely. In other words, the excess return should be calculated based on the total risk (including systematic and non-systematic risks). The Sharpe index (SI), applying the concept of the capital market line,[3] can be written as:

$$SI = \frac{\overline{r}_p - r_f}{\sigma_p}, \tag{22.2}$$

where σ_p is the standard deviation of the portfolio, namely total risk. The Sharpe index is expressed as the reward per unit of total risk. The higher the two indices mentioned above, the better the fund's performance. Because this measure is based on the total risk, it enables to measure the performance of the portfolio which is not very diversified.

Jensen (1968) proposes a regression-based view to measure the performance of the portfolio. The Jensen index (also called Jensen alpha) utilizes the CAPM to determine whether a fund manager outperformed the market. The Jensen index is written as follows:

$$R_{p,t} = \alpha_p + \beta_{p,t} R_{m,t} + u_{p,t}, \tag{22.3}$$

where $R_{p,t}$ and $R_{m,t}$ are the excess returns ($R_t = r_t - r_f$) of the mutual fund and the market portfolio at time t respectively. The term $u_{p,t}$ in the formula is the residual at time t. The coefficient α_p is used to measure the performance of mutual funds for the additional return due to the manager's choice. It also represents the fund manager's selectivity ability without considering the timing ability. A significantly positive and high value of Jensen alpha indicates superior performance compared with the market index.

It should be noted that all three performance measures are interrelated. If the portfolio is well-diversified, then $\rho_{pm} (= \sigma_{pm}/\sigma_p \sigma_m)$ is very close to 1. The Jensen index divided by σ_p can become equivalent to the combinations of Sharpe indices. Since $\beta_p = \sigma_{pm}/\sigma_m^2$, the Jensen index must be multiplied by $1/\sigma_p$ in order to derive the equivalent Sharpe index:

$$\frac{JI}{\sigma_p} = \frac{\overline{r}_p - r_f}{\sigma_p} - \frac{\overline{r}_m - r_f}{\sigma_m}\frac{\sigma_{pm}}{\sigma_m \sigma_p} = \frac{\overline{r}_p - r_f}{\sigma_p} - \frac{\overline{r}_m - r_f}{\sigma_m} = SI_P - SI_m.$$

If the Jensen index is divided by β_P, it is equivalent to the Treynor index with some constant:

$$\frac{JI}{\beta_p} = \frac{\overline{r}_p - r_f}{\beta_p} - \frac{\beta_p(\overline{r}_m - r_f)}{\beta_p} = TI_p - (\overline{r}_m - r_f) = TI_p - \text{constant}.$$

[3] In the presence of a risky asset, this straight line is the efficient frontier for all investors.

The types for the Treynor index and the Sharpe index are very similar. Based on a well-diversified portfolio, the $\beta_p = \rho_{pm}\sigma_p/\sigma_m$ can be simply replaced by σ_p/σ_m. Then the Treynor index can be written as:

$$TI = \sigma_m \frac{\overline{r}_p - r_f}{\sigma_p} = \sigma_m SI.$$

The measures for the Treynor index and Jensen alpha have the same drawback pointed by Roll (1977),[4] the reference index. In addition, when considering a market timing strategy with the dynamic beta changed by forecasting the market movements, the Jensen alpha often becomes negative[5] and does not reflect the true performance of the manager. We present the following methods by allowing the beta variations.

Treynor and Mazuy (1966), putting a quadratic term of the excess market return into (22.3), provide us with a better framework for the adjustment of the portfolio's beta to test a fund manager's timing ability. The fund manager with timing ability will be able to adjust the risk exposure from the market. To take a simple example, if a fund manager expects a rising (down) market, he will hold a larger (smaller) proportion of the market portfolio. Therefore, the portfolio return can be viewed as a convex function of the market return and is given as follows:

$$R_{p,t} = \alpha_p + \beta_1 R_{m,t} + \beta_2 R_{m,t}^2 + \varepsilon_{p,t}, \tag{22.4}$$

where the coefficient β_2 is used to measure the timing ability. When β_2 is significantly larger than zero, it represents that, in an up market, the increasing proportion in the risk premium of the mutual fund is larger than that in the market portfolio. This model was formulated empirically by Treynor and Mazuy (1966). It was then theoretically validated by Jensen (1972) and Bhattacharya and Pfleiderer (1983).

Henriksson and Merton (1981) used the options theory to explain the timing ability. It consists of a modified version of the CAPM which takes the manager's two objectives into account, and depends on whether he forecasts that the market return will or will not be better than the risk-free asset return. They view the coefficient β as a binary variable. This means that a fund manager with market timing ability should have different β values in the up and down markets. We can express the equation as:

$$R_{p,t} = \alpha_p + \beta_1 R_{m,t} + \beta_2 Max(0, -R_{m,t}) + \varepsilon_{p,t}. \tag{22.5}$$

For an up market ($R_{m,t} = r_m - r_f > 0, Max(0, -R_m) = 0$), equation (22.5) can be expressed as $R_{p,t} = \alpha_p + \beta_1 R_{m,t} + \varepsilon_{p,t}$, and for a down market, it can be written

[4] The core criticism relates to the fact that it is impossible to get the true market portfolio. However, Stambaugh (1982) showed that the CAPM is not very sensitive to the market portfolio setting.

[5] Grant (1977) showed that the market timing ability will cause the selectivity ability to be downward biased. However, this point is not consistent with Jensen (1968).

as $R_{p,t} = \alpha_p + (\beta_1 - \beta_2)R_{m,t} + \varepsilon_{p,t}$. If the manager has market timing ability, that's $\beta_2 > 0$, it means that he will hold lower proportion of the market portfolio.

Jensen (1972) showed that the timing ability can be measured by the correlation between the managers' forecast and the realized return. Bhattacharya and Pfleiderer(1983) modified Jensen (1972) model[6] to propose a regression-based model to evaluate the market timing and selectivity abilities. Assume $\pi_t = R_{m,t} - E(R_m)$, where $E(R_m)$ is the unconditional mean of $R_{m,t}$. Moreover, we assume π_t^* as the conditional expected value of π_t, expressed as $E(\pi_t|\phi_t)$ that refers the expected value of π_t under ϕ_t, the manager's information set prior to time t. Then the relationship between π_t^* and π_t is $\pi_t^* = \psi(\pi_t + \varepsilon_t)$. For minimizing the variance of the forecasting error, we determine the optimal value[7] of $\psi = \sigma_\pi^2/(\sigma_\pi^2 + \sigma_\varepsilon^2)$, representing the ratio of the forecasted change to the realized change in the market information. This ratio also reflects their relevance.

In the market equilibrium, the systematic risk coefficient $\beta_{p,t}$ is the same as the target risk coefficient $\beta_{p,T}$. However, if the fund manager owns other specific information, both π_t and π_t^* will be different. Then, the systematic risk $\beta_{p,t}$ of the portfolio will be affected by both π_t^* and the fund manager's sensitivity θ with specific information, ϕ_t. The relationship can be expressed as $\beta_{p,t} = \beta_{p,T} + \theta\pi_t^*$. Suppose the target systematic risk of mutual funds is decided by the reaction for the net market return, i.e., $\beta_{p,T} = \theta E(R_m)$, we can rewrite the Jensen (1972) model (22.3) as:

$$R_{p,t} = \alpha_p + \theta\{E(R_m) + \psi[R_{m,t} - E(R_m) + \varepsilon_t]\}(R_{m,t}) + u_{p,t}. \quad (22.6)$$

Rearranging equation (22.6), we get:

$$R_{p,t} = \alpha_p + \theta E(R_m)(1-\psi)R_{m,t} + \psi\theta(R_{m,t})^2 + \theta\psi\varepsilon_t R_{m,t} + u_{p,t}. \quad (22.7)$$

It also can be written as:

$$R_{p,t} = \eta_0 + \eta_1 R_{m,t} + \eta_2(R_{m,t})^2 + \omega_t. \quad (22.8)$$

[6] In a framework of Jensen (1972), the coefficients in the model can not be estimated efficiently. However, with some assumptions proposed by Bhattacharya and Pfleiderer (1983), we can get the efficient estimators. The detail can be found in Lee and Rahman (1990), pp. 265–266.

[7]
$$\begin{aligned}
&\min_\psi E[\pi_t - \psi(\pi_t + \varepsilon_t)]^2\\
&= \min_\psi E[\pi_t^2 - 2\pi_t\psi(\pi_t + \varepsilon_t) + \psi^2(\pi_t + \varepsilon_t)^2]\\
&= \min_\psi E(\pi_t^2 - 2\pi_t^2\psi - 2\pi_t\varepsilon_t\psi + \psi^2\pi_t^2 + 2\psi^2\pi_t\varepsilon_t + \psi^2\varepsilon_t^2)\\
&= \min_\psi E(\pi_t^2 - 2\pi_t^2\psi + \psi^2\pi_t^2 + \psi^2\varepsilon_t^2)\ (\because E(\pi_t) = E(\varepsilon_t) = E(\pi_t\varepsilon_t) = 0)\\
&= \min_\psi E((1-\psi)^2\pi_t^2 + \psi^2\varepsilon_t^2)\\
&= \min_\psi (1-\psi)^2\sigma_\pi^2 + \psi^2\sigma_\varepsilon^2\\
&\text{FOC:} -2(1-\psi)\sigma_\pi^2 + 2\psi\sigma_\varepsilon^2 = 0,\ \psi = \sigma_\pi^2/(\sigma_\pi^2 + \sigma_\varepsilon^2).
\end{aligned}$$

In a large sample setting, the estimated coefficients are:

$$\operatorname{plim} \eta_0 = \alpha_p, \quad \operatorname{plim} \eta_1 = \theta E(R_m)(1-\psi), \quad \operatorname{plim} \eta_2 = \theta\psi. \tag{22.9}$$

Bhattacharya and Pfleiderer (1983) also use the term α_p to evaluate the selectivity ability, but use the correlation ψ $\left(\rho^2 = \psi = \sigma_\pi^2 \big/ (\sigma_\pi^2 + \sigma_\varepsilon^2)\right)$ between the forecasted and real values of the excess market return to measure the market timing ability.[8] In short, ψ is determined by σ_ε^2 and σ_π^2. The first term, σ_ε^2, can be calculated from ω_t, i.e., $\omega_t = \theta\psi\varepsilon_t R_{m,t} + u_{p,t}$. We can regress ω_t^2 on $R_{m,t}^2$,

$$\omega_t^2 = \theta^2\psi^2\sigma_\varepsilon^2 (R_{m,t})^2 + \tau_t, \tag{22.10}$$

where $\tau_t = \theta^2\psi^2(R_{m,t})^2(\varepsilon_t^2 - \sigma_\varepsilon^2) + (u_{p,t})^2 + 2\theta\psi(R_{m,t})\varepsilon_t u_{p,t}$. Therefore, we can derive $\sigma_\varepsilon^2 = \dfrac{\theta^2\psi^2\sigma_\varepsilon^2}{\theta^2\psi^2} = \dfrac{\theta^2\psi^2\sigma_\varepsilon^2}{\eta_2^2}$.

Additionally, in the assumption of the stationary Wiener process of π_t, Merton (1981) proposes a simple estimator for σ_π^2 without estimating the expected return in advance, i.e., $\sigma_\pi^2 = \frac{1}{n}\sum_{t=1}^{n} [\ln(1 + R_{m,t})]^2$. When ρ (always positive) is significantly different from zero, it means that the fund manager can forecast the market trend. Lee and Rahman (1990) find that the residual terms of the above two equations (22.8) and (22.10) exist heteroscedasticity. This shows that the coefficients, η_0, $\theta^2\psi^2\sigma_\varepsilon^2$, and η_2, estimated from OLS are not efficient. This problem can be solved by calculating the variance of the residuals, ω_t and τ_t. They can be shown as:

$$\sigma_\omega^2 = \theta^2\psi^2\sigma_\varepsilon^2(R_{m,t})^2 + \sigma_u^2, \tag{22.11}$$

$$\sigma_\tau^2 = 2\theta^4\psi^4\sigma_\varepsilon^4 + 2\sigma_u^4 + 4\theta^2\psi^2\sigma_\varepsilon^2(R_{m,t})^2\sigma_u^2, \tag{22.12}$$

where σ_u^2 is the variance of the residual u, as defined in equation (22.3). Using σ_ε^2 and σ_π^2 calculated earlier, we can get the variances σ_ω^2 and σ_τ^2. In order to estimate efficiently, we utilize a GLS method with correction for heteroscedasticity to adjust the weights of the variables in equation (22.8) and (22.10) by σ_ω^2 and σ_τ^2.

22.3 Empirical Results

We utilize different methods to examine the selectivity, market timing, and overall performance for the open-end equity mutual funds.[9] The samples used in this

[8] The negative timing ability is ignored in this model. Bhattacharya and Pfleiderer (1983) suppose the negative correlation still shows the fund manager owns the timing ability, and just mistakes in information applications.

[9] In addition to equity funds (the code in CRSP is EQ), Standard & Poor's Main Category provides the other four kinds of funds to define fund styles, i.e., fixed income (FI), money market (DG), asset allocation (AA), and convertible (CT), but they are not analyzed in this study.

study were the monthly returns of the 628 mutual funds[10] from January 1990 to September 2005; 189 monthly observations. The fund data were obtained from the CRSP Survivor-Bias-Free US Mutual Fund Database. Subsequently, we used the ICDI's fund objective codes to sort the objectives of the mutual funds. In total, there are 23 types of mutual fund objectives. To simplify the empirical process, we divide them as two groups, growth funds and non-growth funds.[11] Finally, our empirical study consists of 439 growth funds and 189 non-growth funds. The mean (the standard deviation) for the beta values of growth funds and non-growth funds are 0.9696 (0.1881) and 0.7673 (0.2947) respectively.

In addition to the CRSP fund data, the S&P 500 stock index obtained from Datastream is used for the return of the market portfolio. Moreover, we use the Treasure bill rate with a 3-month holding period as the risk-free return. The Treasury bill rate is available from the website of the Federal Reserve Board.

Figure 22.1 shows the relationship between the growth and non-growth fund excess returns (net of risk-free return), and the market excess return from January 1990 to September 2005. The fund return process is built by the equal weighted average of the funds in the same group. However, the difference of the scatter plots is small for both the growth (Fig. 22.1a) and non-growth funds (Fig. 22.1b). Both of them cannot display a convex relationship between the fund excess return and market excess return. From the previous explanation, a convex relationship implies that the fund has the market timing ability. Obviously, we need to take a better look at the performance of the individual funds in the growth and non-growth funds. For the overall period, the mean and the standard deviation of the growth fund monthly returns on average are 0.0082 and 0.0431. For the non-growth funds, those are 0.0072 and 0.0373 respectively. Moreover, those for the market are 0.0066 and 0.0416. The means of the two groups are larger than the market.

Figure 22.2 plots the graphs for the market index (S&P 500) and risk-free annualized rate (T-bill rate). It is interesting to compare the fund performance for the different market situations. We divide the entire sample into two subperiods by the end of 1997. As shown in Fig. 22.2a, the former subperiod represents an obvious upward tendency, while the latter does not show any clear trend. The fund performance between these distinct samples will give us a more informative and meaningful inference for our empirical study.

In the following empirical examinations, the fund performances are tested by the Treynor index, Sharpe index, Jensen index, Treynor-Mazuy model, Henriksson-Merton model, and Lee-Rahman model. The tests are run for the entire period, January 1990 to September 2005, and the two subperiods January 1990 to December 1997 and November 1997 to September 2005.

[10] We delete the funds with any missing values in this period. In addition, the funds with over 20 zero returns are also dropped. This is because we view too many zero values as missing data or lower liquidity for the fund. The list of the fund names is available from the authors on request.

[11] Three types of mutual funds belongs to the growth group, i.e., aggressive growth (AG), growth and income (GI), and long-term growth (LG); the other 20 types are put in the non-growth group.

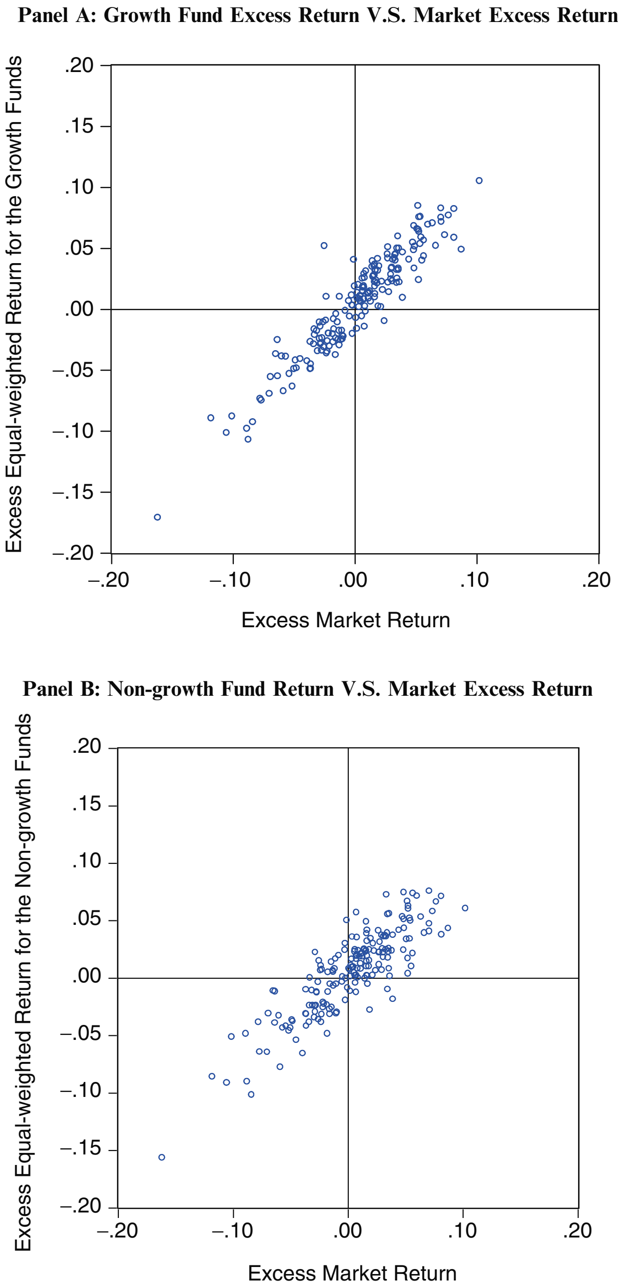
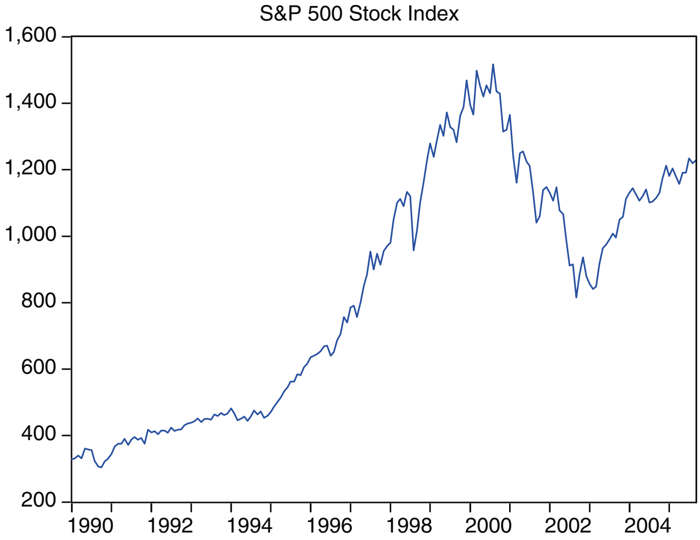


Fig. 22.1 The scatter between fund and market excess returns. This figure shows the relationship between the growth (**a**) and non-growth fund excess returns (**b**) (net of risk-free rate), and the market excess return (S&P 500 index) from January 1990 to September 2005. The fund return process is built by the equal weighted average of the funds in the same group

Panel A: S&P 500 stock index (settlement price)



Panel B: Treasure bill rate (annualized)

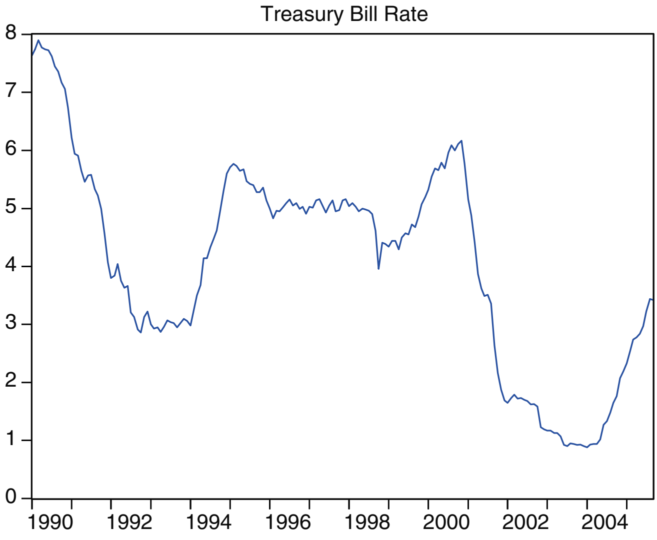


Fig. 22.2 S&P 500 stock index and treasure bill rate. This figure shows the graphs for the market index (S&P 500) (**a**) and risk-free rate (T-bill rate) (**b**) in this study. The sample period is from January 1990 to September 2005

Table 22.1 presents the mutual fund performance measured by the Treynor index. The result for the entire period shows that 82% (69%) of the growth (non-growth) funds have better performance than the market. This seems to point that the growth

Table 22.1 Mutual fund performance measured by Treynor index

Objective	Mean	Std. dev.	Max	Min	Over the market
1990:01–2005:09		Market Treynor index = 0.0032			
Growth	0.0052	0.0025	0.0193	−0.0024	360 (82.0%)
Non-growth	0.0045	0.0070	0.0221	−0.0609	131 (69.3%)
1990:01–1997:10		Market Treynor index = 0.0061			
Growth	0.0082	0.0023	0.0230	−0.0039	388 (88.4%)
Non-growth	0.0027	0.0196	0.1144	−0.1210	110 (58.2%)
1997:11–2005:09		Market Treynor index = 0.0004			
Growth	0.0024	0.0034	0.0154	−0.0082	302 (68.8%)
Non-growth	0.0053	0.0059	0.0291	−0.0108	167 (88.4%)

This table shows the performance of the growth and non-growth mutual funds by Treynor index for three different sample periods. The values in the parentheses represent the ratio of the funds which perform better than the market. The entire sample period is from January 1990 to September 2005, totally, 189 monthly observations

Table 22.2 Mutual fund performance measured by Sharpe index

Objective	Mean	Std. Dev.	Max	Min	Over the market
1990:01–2005:09		Market Sharpe index = 0.0774			
Growth	0.1037	0.0439	0.2695	−0.0455	326 (74.3%)
Non-growth	0.0824	0.0606	0.2860	−0.1296	115 (60.8%)
1990:01–1997:10		Market Sharpe index = 0.1712			
Growth	0.1976	0.0505	0.3631	−0.0327	321 (73.1%)
Non-growth	0.1148	0.1116	0.3933	−0.2355	67 (35.4%)
1997:11–2005:09		Market Sharpe index = 0.0088			
Growth	0.0392	0.0546	0.1958	−0.1429	300 (68.3%)
Non-growth	0.0621	0.0544	0.2644	−0.0993	164 (86.8%)

This table shows the performance of the growth and non-growth mutual funds measured by Sharpe index for three different sample periods. The values in the parentheses represent the ratio of the funds which perform better than the market. The entire sample period is from January 1990 to September 2005, totally, 189 monthly observations

funds are more valuable to be invested than the non-growth funds.[12] For the subperiod 1990–1997, the apparent bull market strengthens the growth funds' performance, but weakens the non-growth funds' performance. Compared with those in the subperiod 1997–2005, the market is not clear. However, the result is very different from the previous findings. The non-growth funds even have more outstanding performance than the growth funds. Clearly, the results indicate that the non-bull market has a negative effect on the growth funds.

The performance of the Treynor index is mainly based on the systematic risk obtained from the CAPM. It is appropriate to evaluate the well-diversified portfolio. In contrast to the Treynor index, the Sharpe index takes account of the total risk of the portfolio. The performance measured by the Sharpe index is given in Table 22.2.

[12] We do not consider the transaction costs and taxes here. In general, the growth fund will ask for a higher commission than the non-growth fund.

Table 22.3 Mutual fund performance measured by Jensen index

Objective	Mean	Std. dev.	Max	Min	$\alpha_p > 0$	$\alpha_p < 0$
1990:01–2005:09						
Growth	0.0017	0.0020	0.0080	−0.0069	360 (108)	79 (1)
Non-growth	0.0013	0.0029	0.0087	−0.0118	131 (22)	58 (1)
1990:01–1997:10						
Growth	0.0020	0.0020	0.0101	−0.0104	388 (92)	51 (1)
Non-growth	−0.0002	0.0043	0.0117	0.0220	109 (16)	80 (2)
1997:11–2005:09						
Growth	0.0015	0.0029	0.0084	−0.0125	302 (48)	137 (7)
Non-growth	0.0030	0.0031	0.0104	−0.0055	167 (10)	22 (0)

This table provides the performance of the growth and non-growth mutual funds measured by Jensen index for three different sample periods. The values in the parentheses represent the number of funds with significant interaction at the 95% level. The entire sample period is from January 1990 to September 2005, totally, 189 monthly observations

Generally speaking, the result is consistent with that in Table 22.1. Except for the subperiod 1990–1997, the fund number over the market is slightly less than the previous one.

Unlike the first two measures, the Jensen index is calculated by carrying out a simple regression. The Jensen alpha is also based on the CAPM and measures the share of additional return that is due to the manager's choices. Table 22.3 summarizes the result of the Jensen index for three periods. Most of the funds have positive Jensen alphas, especially for the growth funds of the subperiod 1990–1997 (88%) and the non-growth funds of the subperiod 1997–2005 (88%). As for the significance of the coefficient alpha, it depends on the investment length and the market trend. Thirty percent of the growth funds for the entire period are significantly larger than zero with 95% confidence. Also, the growth funds of the subperiod 1997–2005 have lower proportion than the non-growth funds, and seven funds have significantly negative alpha values. The mean of Jensen index for the growth funds of the subperiod 1990–1997 is negative. This can be explained by their poor performance relative to the market index for Treynor and Sharpe indices.

The first three indicators assume that the portfolio risk is stationary and only take the stock selection into account. If we want to modify the level of the portfolio's exposure to the market risk, the timing ability has to be adopted. The models with the ability to test the market timing are against the CAPM. This is due to permit variations in the portfolio's beta over the investment period. There are three models for evaluating the selectivity and timing abilities in this study.

Table 22.4 provides the performance of the mutual funds, as measured by the Treynor-Mazuy's model, a quadratic version of the CAPM. The growth funds in the subperiod 1990–1997 have little evidence of the timing ability; only 16 of them have significantly positive estimates with 95% confidence. Except for these, the other results show no timing ability for nearly all funds. In fact, over 80% of them have negative values of timing ability and a considerable ratio among them have significantly negative estimates. Moreover, no one exhibits significantly positive estimates in both the subperiods.

Table 22.4 Mutual fund performance measured by the Treynor-Mazuy's model

Panel A: Selectivity ability

Objective	Mean	Std. dev.	Max	Min	$\alpha_p > 0$	$\alpha_p < 0$
1990:01–2005:09						
Growth	0.0035	0.0029	0.0116	−0.0057	402 (169)	37 (0)
Non-growth	0.0042	0.0029	0.0114	−0.0084	182 (49)	7 (0)
1990:01–1997:10						
Growth	0.0027	0.0030	0.0130	−0.0163	375 (111)	64 (1)
Non-growth	0.0023	0.0042	0.0131	−0.0182	152 (28)	37 (0)
1997:11–2005:09						
Growth	0.0042	0.0042	0.0157	−0.0129	376 (96)	63 (4)
Non-growth	0.0066	0.0054	0.0218	−0.0039	184 (44)	5 (0)

Panel B: Timing ability

Objective	Mean	Std. dev.	Max	Min	$\beta_2 > 0$	$\beta_2 < 0$
1990:01–2005:09						
Growth	−1.0379	1.1441	1.2844	−5.4791	68 (1)	371 (151)
Non-growth	−1.6293	1.4283	1.0747	−6.3699	23 (0)	166 (71)
1990:01–1997:10						
Growth	−0.6109	1.5554	4.6919	−6.1783	160 (21)	279 (56)
Non-growth	−1.9597	1.9150	3.7024	−7.9143	23 (1)	166 (50)
1997:11–2005:09						
Growth	−1.2081	1.2509	1.5759	−6.5961	60 (1)	379 (90)
Non-growth	−1.6374	1.7949	1.4023	−7.7479	28 (0)	161 (40)

This table provides the selectivity and timing abilities of the growth and non-growth mutual funds measured by the Henriksson-Merton's model for three different sample periods. The value in the parentheses represents the number of funds with significant interaction at the 95% level. The entire sample period is from January 1990 to September 2005, totally, 189 monthly observations

For the selectivity ability, a very high ratio of the funds has positive estimates and many of them are significantly positive. Only very few funds have significantly negative estimates. None of the funds have significantly positive or negative estimates of selectivity or timing ability in both periods.

The correlations between the estimates of timing and selectivity ability are t0.62 for the entire period, −0.44 for the subperiod 1990–1997, and −0.77 for the subperiod 1997–2005. Compared with Table 22.3 that considers the timing ability, the estimates for the selectivity ability have higher values. This is also consistent with Grant (1977) and Lee and Rahman (1990).

The Henriksson-Merton model is also a modified version of the CAPM which applies a binary choice for the manager. It depends on whether the market returns will or will not perform better than the risk-free asset return. The results are shown in Table 22.5. In general, the results are similar to those in Table 22.4. The estimates of the selectivity ability even display larger values. Furthermore, the correlations between the estimates of timing and selectivity ability are -0.85 for the entire period, -0.76 for the subperiod 1990–1997, and -0.90 for the subperiod 1997–2005. They show a more substantial relation than the previous one.

Table 22.5 Mutual fund performance measured by the Henriksson-Merton's model

Panel A: Selectivity ability

Objective	Mean	Std. dev.	Max	Min	$\alpha_p > 0$	$\alpha_p < 0$
1990:01–2005:09						
Growth	0.0049	0.0042	0.0173	−0.0108	404 (152)	35 (0)
Non-growth	0.0069	0.0043	0.0228	−0.0062	186 (61)	3 (0)
1990:01–1997:10						
Growth	0.0035	0.0045	0.0152	−0.0280	356 (102)	83 (1)
Non-growth	0.0052	0.0051	0.0196	−0.0141	170 (42)	19 (0)
1997:11–2005:09						
Growth	0.0060	0.0058	0.0212	−0.0091	381 (82)	58 (0)
Non-growth	0.0094	0.0085	0.0382	−0.0054	175 (44)	14 (0)

Panel B: Timing ability

Objective	Mean	Std. dev.	Max	Min	$\beta_2 > 0$	$\beta_2 < 0$
1990:01–2005:09						
Growth	−0.1956	0.2283	0.3341	−1.1735	75 (2)	364 (110)
Non-growth	−0.3435	0.3063	0.2302	−1.4967	23 (0)	166 (67)
1990:01–1997:10						
Growth	−0.1131	0.2563	1.2541	−0.9976	156 (13)	283 (45)
Non-growth	−0.3807	0.3782	0.7033	−1.6574	22 (1)	167 (55)
1997:11–2005:09						
Growth	−0.2442	0.2729	0.4193	−1.3411	74 (0)	365 (49)
Non-growth	−0.3494	0.4187	0.4233	−1.9448	36 (0)	153 (39)

This table provides the selectivity and timing abilities of the growth and non-growth mutual funds measured by the Henriksson-Merton's model for three different sample periods. The value in the parentheses represents the number of funds with significant interaction at the 95% level. The entire sample period is from January 1990 to September 2005, totally, 189 monthly observations

Table 22.6 provides the performance of the funds, as measured by the Lee-Rahman model. The result for the selectivity ability is very similar to that given in Table 22.4. Because the Lee-Rahman model assumes no negative timing ability, it is worth discussing its results. The non-growth funds have better timing ability than the growth funds for all periods. Different from the previous two models, the correlations between the estimates of timing and selectivity ability are 0.43 for the entire period, 0.24 for the subperiod 1990–1997, and 0.45 for the subperiod 1997–2005. All of them are positive. Moreover, 40 of the funds in the entire period have significantly positive estimates in both selectivity and timing ability. Nevertheless, none of the funds have significantly positive estimates in both selectivity and timing ability in both subperiods.

How about the performance of the funds with significantly positive selectivity and timing ability? Before considering it, we show the actual return with an initial investment of \$1.0 at the beginning of the sample period for the mutual funds, the market portfolio, and the risk-free asset, as shown in Table 22.7 (see footnote 12).

Table 22.7 shows that the growth and non-growth funds have very different performance with the market. In the subperiod 1990–1997, 83% of the growth funds

Table 22.6 Mutual fund performance measured by the Lee-Rahman's model

Panel A: Selectivity ability

Objective	Mean	Std. dev.	Max	Min	$\alpha_p > 0$	$\alpha_p < 0$
1990:01–2005:09						
Growth	0.0033	0.0028	0.0109	−0.0062	399 (150)	40 (0)
Non-growth	0.0041	0.0026	0.0101	−0.0080	183 (48)	6 (0)
1990:01–1997:10						
Growth	0.0028	0.0030	0.0133	−0.0165	377 (123)	62 (1)
Non-growth	0.0029	0.0042	0.0133	−0.0180	156 (34)	33 (0)
1997:11–2005:09						
Growth	0.0038	0.0039	0.0156	−0.0086	361 (77)	78 (0)
Non-growth	0.0061	0.0050	0.0201	−0.0048	178 (37)	11 (0)

Panel B: Timing ability

Objective	Mean	Std. dev.	Max	Min	ρ
1990:01–2005:09					
Growth	0.0816	0.0510	0.2274	0.0000	439 (36)
Non-growth	0.0982	0.0621	0.2409	0.0069	189 (41)
1990:01–1997:10					
Growth	0.1086	0.0732	0.4081	0.0007	439 (24)
Non-growth	0.1442	0.0745	0.2896	0.0034	189 (19)
1997:11–2005:09					
Growth	0.0997	0.0612	0.2599	0.0002	439 (7)
Non-growth	0.1118	0.0672	0.2975	0.0004	189 (12)

This table provides the performance of the growth and non-growth mutual funds measured by the Lee-Rahman's model for three different sample periods. The value in the parentheses represents the number of funds with significant interaction at the 95% level. The correlation ρ between the managers' forecast and the realized return is always positive. The entire sample period is from January 1990 to September 2005, totally, 189 monthly observations

performed better than the market, but only 38% of the non-growth funds did so. In the subperiod 1997–2005, 69% of the growth funds still performed better than the market. However, for the non-growth funds, 86% of them performed better than the market. On an average, for the initial investment of \$1.0 at the beginning of 1990, the growth funds and the non-growth funds will get \$5.0 and \$4.6 in the end, respectively. Both of them are more than the market (\$3.5) and the risk-free asset (\$1.9). Figure 22.3 highlights the difference between the growth and non-growth funds for the three periods.

How about the 40 funds with significantly positive estimates in the selectivity and timing ability? The mean of the final amount for them is only \$5.1, very close to the overall mean. From our empirical study, we seem to be able to conclude that the selectivity and timing abilities are not the key factors to decide the fund's performance.

Table 22.7 Return on the initial investment of $1.00

	1990:01–2005:09		1990:01–1997:10		1997:11–2005:09	
	Growth	Non-growth	Growth	Non-growth	Growth	Non-growth
Mean	4.9902	4.5626	3.0965	2.5022	1.6080	1.8503
Std. Dev.	1.7289	2.5958	0.6060	1.3274	0.4373	0.5418
Max	13.1251	14.4838	5.4613	8.4894	2.9991	3.5658
Min	1.0856	0.2307	1.0115	0.1758	0.4177	0.7859
Market portfolio	3.4771		2.5881		1.3435	
Risk-free securities	1.8921		1.4646		1.2919	
Fund return greater than market portfolio	362 (82.5%)	119 (63.0%)	363 (82.7%)	71 (37.6%)	304 (69.2%)	163 (86.2%)
Fund return greater than risk-free securities	436 (99.3%)	174 (92.1%)	437 (99.5%)	155 (82.0%)	331 (75.4%)	168 (88.9%)

This table shows the final returns of the initial investment $1.00 at the beginning of the sample period for the growth and non-growth mutual funds, market portfolio, and risk-free securities. We also compare the performance of mutual funds with market portfolio and risk-free securities for three different sample periods. The values in the parentheses represent the ratio of funds which perform better than the market (the risk-free asset). The entire sample period is from January 1990 to September 2005, totally, 189 monthly observations

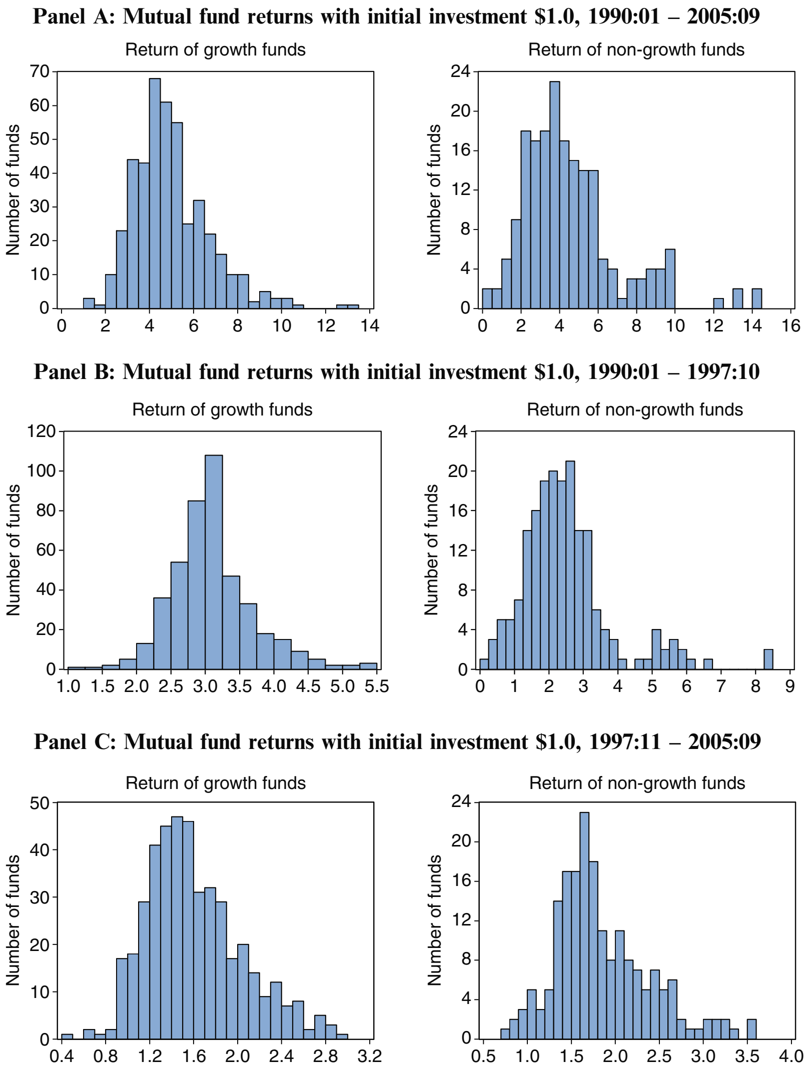


Fig. 22.3 Growth and non-growth mutual fund returns with initial investment of $1.0 for the entire period and two subperiods. This figure shows the growth and non-growth fund returns of the initial investment $1.00 at the beginning for three different sample periods. The entire sample period is from January 1990 to September 2005, totally, 189 monthly observations. There are 439 growth funds and 189 non-growth funds

22.4 Conclusion

This paper empirically examines selectivity, market timing, and overall performance of equity funds in the United States during January 1990 until September 2005. We first used Sharpe, Treynor, and Jensen measures to evaluate the selectivity performance of mutual fund managers. In addition, we also used the Treynor-Mazuy, Henriksson-Merton, and Lee-Rahman models to evaluate the selectivity and timing performance of mutual fund managers. The funds, divided into two groups, growth and non-growth, are examined for the entire period and two subperiods. In addition to testing the difference of the performance for two distinct groups, we want to test whether their performance will vary with the market.

The findings support that the growth funds perform better than the non-growth funds in the overall holding period. However, their performances are easily affected by the market condition. The performance for the real investment also supports this inference. As for the selectivity and timing abilities, about one-third of the funds have the selectivity ability, but very few have the timing ability. Moreover, a fund with both significantly positive selectivity and timing abilities does not guarantee to get a superior performance.

References

Bhattacharya, S., and P. Pfleiderer, "A Note on Performance Evaluation." Technical Report 714, Stanford, CA: Stanford University, Graduate School of Business (1983).

Brinson, G. P., B. D. Singer, and G. L. Beebower, "Determinants of Portfolio Performance II: An Update." Financial Analysts Journal 47, 40–48 (1991).

Fama, E. F., "Components of Investment Performance." Journal of Finance 27, 551–567 (1972).

Grant, D., "Portfolio Performance and the Cost of Timing Decisions." Journal of Finance 32, 837–846 (1977).

Henriksson, R. D., and R. C. Merton, "On Market Timing and Investment Performance. II. Statistical Procedure for Evaluating Forecasting Skills." Journal of Business 54, 513–534 (1981).

Jensen, M.C., "The Performance of Mutual Funds in the Period 1945–1964." Journal of Finance 23, 389–416 (1968).

Jensen, M. C., "Optimal Utilization of Market Forecasts and the Evaluation of Investment Performance." In G. P. Szego and K. Shell(eds.), Mathematical Methods in Investment and Finance. Amsterdam: Elsevier (1972).

Lee C. F., and S. Rahman, "Market Timing, Selectivity, and Mutual Fund Performance: An Empirical Investigation." Journal of Business 63, 261–278 (1990).

Markowitz, H., "Portfolio Selection." Journal of Finance 7, 77–91 (1952).

Merton, R. C., "On Market Timing and Investment performance. I. An Equilibrium Theory of Value for Market Forecasts." Journal of Business 54, 363–406 (1981).

Roll, R., "A Critique of the Asset Pricing Theory's Tests, Part I: On Past and Potential Testability of the Theory." Journal of Financial Economics 4, 126–176 (1977).

Sharpe, W. F., "Mutual Fund Performance." Journal of Business 39, 119–138 (1966).

Stambaugh, R. F., "On the Exclusion of Assets from Tests of the Two-Parameter Model: A Sensitivity Analysis." Journal of Financial Economics 10, 237–268 (1982).

Treynor, J. L., "How to Rate Management of Investment Funds." Harvard Business Review 13, 63–75 (1965).

Treynor, J. L., and K. K. Mazuy, "Can Mutual Funds Outguess the Market?" Harvard Business Review 44, 131–136 (1966).

Chapter 23
Case Closed

Robert A. Haugen and Nardin L. Baker[1]

In 1996 we published a paper (Haugen and Baker 1996) on the commonality in the determinants of the cross-section of stock returns over time and across geographical areas. In our 1996 paper, we attempted to explain the cross-section of stock returns with a simple and comprehensive list of stock and company characteristics. These included measures of risk, stock liquidity, profitability and trends in profitability, cheapness in the stock price, and stock price performance in trailing periods. These characteristics were called "factors" and the multiple regression procedure used to estimate the monthly payoffs to the factors an "expected return factor model".

The first expected return factor model in the finance literature was introduced by Fama and MacBeth (1973). Their theoretically guided model included only a few factors, all related to market risk. The selection of factors in our model was intended to be comprehensive and largely unguided by financial theory. As it turned out, our more comprehensive model was more effective than the theoretically guided model in explaining returns in the cross-section.

This chapter extends the results of the comprehensive model to a considerably longer period of time. In greatly extending the period, we find results that are highly consistent with the results of the original paper. We find a remarkable level of power and stability in the factors that are most influential in determining the structure of stock returns. In addition to its *explanatory* power, we find that the model also has amazing and consistent power in *predicting* stocks with relatively high and relatively low future returns. Moreover, we find that, *after allowing for transactions costs*, an optimized portfolio management strategy using the expected return factor model

R.A. Haugen (✉)
Haugen Custom Financial Systems
e-mail: bobh@frontier.net

N.L. Baker
Chief Investment Officer for Quantitative Equity Management

[1] Haugen is the President of Haugen Custom Financial Systems, which licenses the predictions of the expected return factor model to large institutional clients. Baker is the Chief Investment Officer for Quantitative Equity Management, an institutional investment advisor that has successfully managed assets using the expected return factor model for the past 15 years.

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_23,
© Springer Science+Business Media, LLC 2010

outperforms by a comfortable margin. This is true over the total period and within each of the subperiods.

23.1 Methodology and Data

In a given month, we simultaneously estimate the payoffs to a variety of firm and stock characteristics using a weighted least squares multiple regression procedure of the following form:

$$r_{j,t} = \sum_{i=1}^{n} P_{i,t} F_{i,j,t-1} + \mu_{j,t}, \tag{23.1}$$

where $r_{j,t}$ is the total rate of return to stock j in month t, $P_{i,t}$ is the estimated weighted least squares regression coefficient (payoff) for factor i in month t, $F_{i,j,t-1}$ is the normalized value for factor i for stock j at the end of month $t-1$, n is the number of factors in the expected return factor model, and $\mu_{j,t}$ is the component of total monthly return for stock j in month t unexplained by the set of factors.

In 1963, there were 653 companies with sufficient data to be included in the factor estimation procedure. By 1973, there were more than 3,000 and only the top 3,000 market capitalization companies were used in the procedure thereafter.

For accounting numbers, such as earnings-per-share, we used the month-end date after the report date (if available) or a reporting lag of three months (if the report date is unavailable). However, after 1987, the as-reported set of data files that were actually commercially available in the forecast month, were used to calculate all factor exposures. Thus, "look ahead" bias should not significantly affect our results.

Data for all factors are available during the entire period with the exception that three growth factors are not available until February 1964: Dividend-to-price growth, book-to-price growth, and cash flow-to-price growth. If no factor data is available, the payoff to that factor is set to zero for the month.

23.2 The Most Important Factors Explaining the Cross-Sectional Structure of Stock Returns

We estimate (23.1) in each month in the period 1963–2007.[2] In accordance with the study of Fama and MacBeth, we then compute the average values for the monthly regression coefficients (payoffs) across the entire period. Dividing the mean payoffs by their standard errors, we obtain the t-statistics. All the factors are ranked by

[2] Fifty-seven factors are used in the model. See our original paper for definitions.

the absolute values of their t-scores, and the 12 factors with the largest scores are presented in the first column of Table 23.1.

Table 23.1 t-Statistics on the 12 most significant factors

Period	1963–2007	1963–1972	1973–1982	1983–1992	1993–2002	2003–2007
Residual return	−22.4	−13.7	−15.9	−12.9	−7.2	−2.7
Cash flow-to-price	13.9	6.4	12.7	8.6	4.3	4.1
Earnings-to-price	13.1	4.0	11.4	8.3	5.3	1.9
Return on assets	12.6	6.8	7.5	7.5	4.2	3.3
Residual risk	−11.1	−3.5	−6.7	−8.8	−4.7	−1.9
12-month return	10.8	5.0	5.7	6.9	5.1	1.1
Return on equity	10.2	7.0	3.7	6.2	3.9	1.4
Volatility	−9.0	−2.3	−5.6	−7.1	−4.5	−2.0
Book-to-price	8.9	2.0	6.2	6.7	3.2	3.1
Profit margin	7.8	1.0	4.3	6.0	5.7	1.5
3-Month return	−7.2	−5.1	−6.9	−2.8	−.9	−1.5
Sales-to-price	7.0	1.4	3.9	5.3	3.5	2.8

In each month, from January 1963 through December 2007, the cross-sections of the realized stock returns are regressions on seventy characteristics (factors) of each stock using a weighted least squares procedure. The regression coefficients are averaged and t-statistics are computed. The t-statistics for the fifteen most significant factors over the entire period are displayed in the first column.

The values for the most significant factors are computed as follows:

- Residual return is last month's residual stock return unexplained by the market.
- Cash flow-to-price is the12-month trailing cash flow-per-share divided by the current price.
- Earnings-to-price is the 12-month trailing earnings-per-share divided by the current price.
- Return on assets is the 12-month trailing total income divided by the most recently reported total assets.
- Residual risk is the trailing variance of residual stock return unexplained by market return.
- 12-month return is the total return for the stock over the trailing 12 months.
- Return on equity is the12-month trailing earnings-per-share divided by the most recently reported book equity.
- Volatility is the 24-month trailing volatility of total stock return.
- Book-to-price is the most recently reported book value of equity divided by the current market price.
- Profit margin is the 12-month trailing earnings before interest divided by 12-month trailing sales.
- 3-month return is the total return for the stock over the trailing 3 months.
- Sales-to-price is the 12-month trailing sales-per-share divided by the market price.

Note that these scores are larger than those obtained by Fama and MacBeth, even though the length of the time periods covered by the studies is comparable.[3]

Last month's residual return and the return over the preceding three months have negative predictive power relative to next month's total return. This may be induced by the fact that the market tends to overreact to most information. The overreaction sets up a tendency for the market to reverse itself upon the receipt of the next piece of related information.

Four measures of cheapness: cash-to-price, earnings-to-price, book-to-price, and sales-to-price, all have significant positive payoffs. Measures of cheapness have been frequently found in the past[4] to be associated with relatively high stock returns, so it is not surprising that five measures of cheapness appear here as significant determinants of structure in the cross-section.

It could be argued that including all these measures of cheapness in the regressions would make the methodology prone to multicolinearity. Significant problems associated with multicolinearity should result in instability in the estimated regression coefficients from month-to-month. However, as can be seen in Table 23.1, the mean values for these coefficients are very large relative to their standard errors. Multicolinearity is clearly not a significant problem here.

A comprehensive explanation of the positive signs for the various measures of cheapness can be found in Haugen (2004). To be succinct, we feel that the market overreacts to past record of success and failure on the part of companies, making relatively expensive (growth) stocks too expensive and relatively cheap (value) stocks too inexpensive. After the overreaction, the market tends to correct itself, producing low returns to growth stocks and high returns to value stocks, as the relative profitability of these companies tends to mean-revert faster than expected.

Three measures of current profitability: return on assets, return on equity, and profit margin also appear prominently in Table 23.1. These have not been suggested by other authors as significant determinants of relative returns in the cross-section. All are positively related to future return.

A comprehensive explanation of the positive signs for the various measures of profitability can be found in Haugen (2002). In short, we feel that the market prices stocks with a significant degree of imprecision.

To understand this, assume that "true abnormal profit" is the best possible estimate of the risk-adjusted present value of a firm's future abnormal profits – at least the portion that can be expected to accrue to the firm's stockholders. Assume also that "priced abnormal profit" is that which is reflected in the current stock price. In a strictly efficient market, the two measures of abnormal profit should always be equal. In a less than efficient market, they can be different. The market may assign different prices to stocks that have the same true abnormal profit. It may also assign the same price to stocks with different true abnormal profits. We would expect that

[3] In Fama and French (2008) p. 1668, ad hoc cross-section regressions are used in an attempt to explain the cross-sectional structure of stock returns. They report t-statistics as large as −8.59, but no attempt is made to investigate the out-of-sample predictive power of their regressions.

[4] See, for example, Fama and French (1992).

the true abnormal profit is positively correlated with a firm's current measure of profitability. Given that it is, in a market that prices imprecisely, holding everything else (including the stock price) constant, stocks with higher measures of current profitability should be expected to produce higher future returns.

We also see in Table 23.1 that the two measures of risk,[5] including volatility of total return and variance of residual return[6] have highly significant negative coefficients.[7,8]

Once again, a comprehensive explanation for the negative payoffs to risk can be found in Haugen (2002). Here, in brief, the market overreacts to the past success and failure by business firms, pricing the stocks of successful firms too high and the stocks of unsuccessful firms too low. The expensive stocks of successful firms also tend to have higher volatility of return.[9] The overpricing of expensive stocks overrides the markets risk aversion, and the market is consistently *surprised* to find that these relatively more risky stocks tend to produce relatively lower returns.[10]

Table 23.1 reveals that the stocks that pay dividends that produce higher returns than stocks that do not. This tendency may not be related to issues of market efficiency. During most of the period covered by the study, dividends were taxed at higher rates than capital gains. The market may require higher returns on stocks that pay dividends to overcome their tax disadvantage. Ultimately, interpretation will rest on the magnitude of the payoff to paying dividends.

Finally, we note that the momentum (over the trailing 12 months) seems to be positively related to next month's return. This has been found by others[11] and may be related to the fact that the market *underestimates* the tendency for good (or bad) earnings reports to be followed by others of the same sign.

However, far and away the most interesting feature of Table 23.1 is the consistency of the payoffs within the subperiods. We divide the total 45-year period into the first four decades and the final 5 years. It is interesting to note that the majority

---

[5] It should be noted that, although it fails to make to the top 15 most significant factors, market beta has a negative payoff overall and in each of the sub-periods.

[6] Ang, Hodrick, Xing, and Zhang (2006) have recognized the negative payoff to residual risk.

[7] Market beta was in the array of factors used in the regressions, but it did not make to the list of most significant.

[8] Haugen and Heins (1975) were the first to identify the negative payoff to risk in the US stock market. It should be noted that the working paper for this chapter was first released in 1969.

[9] See Lettau and Wachter (2007) p. 60.

[10] In spite of the fact that market overreaction erases the traces of risk aversion in the cross-section, risk aversion can be clearly seen in longitudinal studies of market behavior. As we shall see below, daily returns to the S&P 500 stock index are related to percentage changes in the implied volatility of the index over the period January 1990–May 2008. The relationship between the two is clearly negative with a coefficient of determination of 47%. Increases in the perceived volatility of the index are associated with declines in its level, as the market lowers the price in order to provide higher future returns to investors in the more volatile future period. It is unlikely that a more important determinant of the daily return to the index than reaction to changes in its perceived risk will ever be found.

[11] See for example Jegadeesh and Titman (1993).

of the payoffs continue to be significant in each of the subperiods. However, it is even more striking to note that *they all continue to have the same sign*.[12]

23.3 Predictive Power of the Expected Return Factor Model

By developing a trailing history of the payoffs to the various factors, one can project an expected payoff for the next month. Thus,

$$E(r_{j,t}) = \sum_{i=1}^{n} E(P_{i,t}) F_{i,j,t-1}, \tag{23.2}$$

where $E(r_{j,t})$ is the expected return for stock j in month t, $E(P_{i,t})$ is the mean of a trailing window of 12 monthly payoffs for factor i at time t, and $\underline{F}_{i,j,t-1}$ is the normalized value for factor i for stock j at time $t-1$. The factor value is computed with data that was actually available at $t-1$.

For any given year in the total period, at the beginning of each month all the stocks in the data base are ranked by their expected returns computed in accordance with (23.2) and formed into deciles, where decile 1 has the lowest expected return and decile 10 the highest. The actual total returns are then computed for each decile on an equally weighted basis. The process is repeated for each month of the year, and the 12 monthly returns are linked. Then the linked returns are regressed on the decile ranking to obtain a line of best-fit through the ten plot points. We then calculate the difference between the end of the line (over rank 10) and the beginning of the line (over rank 1). This process is repeated in each year of the 1963–2007 period. The annual spreads are given in Table 23.2.[13]

[12] In interpreting the magnitude of the t-statistics, it is important to remember that the number of observations used in the 03–07 period is half that used for the decades and is 1/9th of the total period.

[13] It is interesting to see the effects of lagging the fundamental data behind the predictions, and what effect lagging has, depending on the length of the horizon over which the accuracy of the predicted returns are evaluated. In the table below, the evaluation month gets further out as we move across the rows, and the length of the data lag increases as we move down the columns. The numbers in the cells correspond with those in Table 23.2 except for the fact that they are measured over the 45-year period.

Evaluation month

Data lag	1 month out (%)	2 months out (%)	3 months out (%)	4 months out (%)
0 months	30.6	21.4	16.7	15.2
1 month	28.0	23.4	20.0	17.9
2 months	25.4	22.6	19.7	18.1
3 months	25.5	22.6	19.8	17.9

Note that the data lag adversely impacts the accuracy of the predictions when accuracy is evaluated in the next month, but there is not much impact for months further away from the estimation month.

Table 23.2 Spreads for decile lines of best fit for each year

Year	Spread (%)	Year	Spread (%)	Year	Spread (%)	Year	Spread (%)
1963	9.2	1974	30.7	1985	36.6	1996	10.4
1964	12.2	1975	30.9	1986	46.4	1997	46.4
1965	30.0	1976	32.4	1987	26.7	1998	23.8
1966	9.4	1977	24.4	1988	18.5	1999	31.9
1967	49.1	1978	7.8	1989	32.2	2000	44.6
1968	13.8	1979	22.1	1990	33.4	2001	57.4
1969	32.4	1980	27.4	1991	27.7	2002	60.2
1970	43.3	1981	33.7	1992	10.6	2003	−5.5
1971	14.7	1982	48.6	1993	14.0	2004	21.1
1972	29.7	1983	39.1	1994	16.8	2005	12.8
1973	44.4	1984	49.7	1995	14.2	2006	7.5
						2007	29.1

At the beginning of each month, the expected return of each stock is calculated by multiplying the normalized value for its factor exposure by the projected factor payoff for the month. The projected payoff is based on the trailing payoffs for the trailing 12-month period. The factor exposure for each stock is based on information that was available at the beginning of each month. This process is repeated for each of the 12 months of each year. At the beginning of each month, stocks are ranked by their expected return and formed into deciles. The 12 monthly realized rates of return for each decile are then linked to form a yearly return. Yearly decile returns are then regressed on decile ranking. The numbers below show the spreads between the regression lines over decile 10 (highest expected return) and decile 1 (lowest expected return)

The table shows that, with the single exception of 2003, the model has positive predictive power every year. Moreover, there appears to be no tendency for the predictive power to wane with the passage of time. The reader should not be unduly impressed with the magnitudes of the spreads, because there is a high level of turnover within the deciles each month, and the spreads do not account for trading costs. The only significance of Table 23.2 to the conclusion of this paper is that the return (gross of trading costs) definitely tends to increase the raw returns as we move from decile 1 to 10. However, in Sect. 23.5, we shall account for the trading costs in the context of a Markowitz-optimized portfolio management strategy.

Some would argue that the spreads are reflective of the differences in risk between the deciles, with decile 10 being more risky and decile 1 less risky. In Sect. 23.4, we investigate the decile characteristics.

The number to the lower right is the spread where you estimate the model with stale, 3-month old data, then wait three months to invest. Thus, any residual issues associated with look-ahead bias have no effect on our proof.

23.4 The Characteristics of Deciles 1–10

Table 23.3 shows the profile characteristics of the ten deciles. The numbers presented in the table are z-scores (number of standard deviations below or above the mean in a normalized distribution for the population).

As we go from decile 1 through decile 10, the transformation in the character of the deciles is absolutely *stunning*. In terms of the risk associated with returns to the stocks, there can be no doubt that the risk decreases as we go from the lowest expected return deciles to the highest. This, of course, is consistent with our findings that the payoff to risk is consistently negative over the 45-year period of this study. The spreads between the extreme deciles are larger for volatility and residual risk than they are for market beta, indicating that these may be more important to pricing than beta. We also see that the high expected return stocks are larger in terms of market capitalization. The inescapable conclusion here is that the *higher* expected return is associated with *lower* risk.

As we go from decile 1 through decile 10, measures of profitability improve dramatically. High expected return stocks are clearly more profitable. Moreover, looking at the trend in profitability over the trailing 5-year window, the high expected return stocks are becoming even more profitable within this window. A higher fraction of the high expected return stocks also tend to pay dividends.

High expected return stocks also sell at cheaper prices relative to earnings, cash flow, sales, and dividends than their low expected return counterparts.

The returns to higher expected return stocks are also larger over the trailing 6- and 12-month periods. High expected return is associated with trailing momentum in the stock price.

In Table 23.4, we see the differences in the z-scores between deciles 10 and 1 for the entire period and for the five subperiods. Consistent with our findings on the stability of the t-statistics in Table 23.1, the characteristics of high and low expected return stocks is amazingly stable over time. The relative nature of the profiles for high and low expected return holds in every period, saves for growth in profit margin in the final 5-year period. There can be no question that the high expected return decile has a more attractive profile than the low expected return decile, and that this relative attraction continues through the decades.

A new type of investment style is being revealed here. The value and growth styles are well known. Some managers also offer a style that is known as "growth at a reasonable price" (GARP), in which they attempt to invest in stocks with good prospects while maintaining discipline in terms of the prices they are willing to pay. Our results reveal that it is possible to go beyond GARP. It is possible to get "growth at a cheap price"(GACP).[14] Individually few, if any, stocks individually have the

[14] This new investment style might be pronounced "gassip".

Table 23.3 Characteristics of deciles 1–10

Decile	1	2	3	4	5	6	7	8	9	10
Market beta	.32	.18	.09	.03	−.02	−.06	−.09	−.13	−.15	−.17
Volatility	.70	.36	.17	.05	−.05	−.12	−.20	−.27	−.32	−.36
Residual risk	.71	.36	.17	.05	−.05	−.12	−.21	−.27	−.33	−.35
Interest coverage	−.42	−.14	−.04	.02	.05	.08	.10	.11	.12	.14
Market capitalization	−.42	−.21	−.12	−.05	.02	.07	.11	.16	.20	.26
Return on assets	−.76	−.27	−.08	.03	.09	.14	.17	.19	.21	.29
ROA growth	−.19	−.09	−.05	−.03	.00	.02	.04	.07	.10	.14
Return on equity	−.68	−.30	−.12	−.01	.06	.11	.15	.19	.24	.36
ROE growth	−.25	−.12	−.06	−.03	.01	.03	.06	.09	.13	.18
Profit margin	−.63	−.22	−.07	.00	.05	.08	.12	.16	.21	.31
PM growth	−.16	−.05	−.02	−.01	.01	.02	.03	.04	.06	.08
Earnings growth	−.24	−.07	−.01	.02	.04	.04	.04	.05	.06	.07
Dividend?	.32	.44	.52	.57	.62	.66	.70	.73	.75	.75
Earnings-to-price	−.84	−.41	−.21	−.07	.04	.13	.22	.29	.36	.49
Cash flow-to-price	−.70	−.31	−.14	−.03	.06	.12	.17	.22	.26	.36
Book-to-price	−.17	−.08	−.05	−.03	.01	.03	.05	.07	.08	.10
Dividend-to-price	−.04	−.04	−.03	−.03	−.02	−.01	.01	.04	.07	.10
6-month return	−.30	−.18	−.11	−.06	−.02	.03	.07	.12	.18	.27
12-month return	−.59	−.33	−.19	−.10	−.02	.06	.13	.22	.33	.51

At the beginning of each month stocks are ranked by their expected return in accordance with the expected return factor model and formed into deciles, with decile 10 having the highest expected return. The characteristics of the stocks are normalized into z-scores. For each characteristic, the average z-score for decile 10 and decile 1 are computed across all stocks in the decile and across all months in the period 1963–2007.

- Market beta is computed by regressing stock returns on the returns to the S&P 500 Stock Index over trailing 24-month periods.
- Volatility is the standard deviation of return over trailing 24-months.
- Residual risk is the variance of return unexplained by the S&P 500 over trailing 24-months.
- Interest coverage is the ratio of operating income to the total interest expense for the most recent 12-month period.
- Market cap is the market price of the stock multiplied by the total shares outstanding at the beginning of the month.
- Return on assets is the most recently reported operating income to total assets.
- Return on equity is the most recently reported net income to book equity.
- Profit margin is the ratio of total operating income to total sales.
- All growth numbers for ratios are obtained by regressing the quarterly values for the ratios on time for a trailing 5-year period.
- Earnings growth is obtained by regressing the log of quarterly earnings-per-share on time.
- Dividend? is an indicator that takes a value of 1 if a stock pays a dividend and 0 otherwise.
- Earnings-to-price is the ratio of the most recently reported earnings-per-share to the market price of the stock at the beginning of the month.
- Cash-to-price is the ratio of cash flow to this same value for market price.
- Sales-to-price is the most recently reported sales-per-share to the same value for market price.
- Dividend-to-price is the total value of dividends paid over the most recent 4-quarters to the same value for market price.
- 6- and 12-month returns are based on total returns for the stock over trailing 6- and 12-month periods.

(Normalized z-scores for deciles)

Table 23.4 Differences in the characteristics of high and low expected return stocks

	1963–1907	1963–1972	1973–1982	1983–1992	1993–1902	2003–1907
Market beta	−.49	−.22	−.70	−.44	−.54	−.62
Volatility	−1.06	−.39	−1.33	−1.49	−1.17	−.78
Residual risk	−1.06	−.37	−1.31	−1.57	−1.17	−.78
Int. coverage	.56	.10	.42	.97	.80	.48
Market cap.	.67	.19	.64	1.17	.77	.50
Return on assets	1.04	.95	.90	1.33	1.07	.88
ROA growth	.33	.44	.34	.53	.08	.15
Return on equity	1.05	.82	.89	1.37	1.22	.84
ROE growth	.42	.47	.36	.56	.36	.32
Profit margin	.94	.35	.98	1.28	1.27	.75
PM growth	.24	.31	.18	.47	.13	−.03
Earnings growth	.31	.16	.12	.45	.43	.44
Dividend?	.43	.27	.51	.59	.41	.29
Earnings-to-price	1.33	.99	1.47	1.62	1.33	1.16
Cash flow-to-price	1.05	.61	1.18	1.24	1.08	1.26
Book-to price	.27	−.02	.58	.31	.13	.41
Dividend-to-price	.14	.05	.38	.11	.04	.06
6-month return	.57	.56	−.03	.85	1.06	.27
12-month return	1.10	1.12	.63	1.34	1.49	.76

At the beginning of each month, stocks are ranked by their expected return in accordance with the expected return factor model and formed into equally weighted deciles, with decile 10 having the highest expected return. The characteristics of the stocks are normalized into z-scores. For each characteristic, the average z-score for decile 10 and decile 1 are computed across all stocks in the decile and across all months in the various periods. The differences between the grand average scores for deciles 10 and 1 are shown in the table for the whole period and for subperiods.

- Market beta is computed by regressing the stock returns on the returns to the S&P 500 Stock Index over trailing 24-month periods.
- Volatility is the standard deviation of return over trailing 24-months.
- Residual risk is the variance of return unexplained by the S&P 500 over trailing 24-months.
- Interest coverage is the ratio of operating income to total interest expense for the most recent 12-month period.
- Market cap is the market price of the stock multiplied by the total shares outstanding at the beginning of the month.
- Return on assets is the most recently reported operating income to total assets.
- Return on equity is the most recently reported net income to book equity.
- Profit margin is the ratio of total operating income to total sales.
- All growth numbers for ratios are obtained by regressing the quarterly values for the ratios on time for a trailing 5-year period.
- Earnings growth is obtained by regressing the log of quarterly earnings-per-share on time.
- Dividend? is an indicator that takes a value of 1 if a stock pays a dividend and 0 otherwise.
- Earnings-to-price is the ratio of the most recently reported earnings-per-share to the market price of the stock at the beginning of the month.
- Cash-to-price is the ratio of cash flow to this same value for market price.
- Sales-to-price is the most recently reported sales-per-share to the same value for market price.
- Dividend-to-price is the total value of dividends paid over the most recent 4-quarters to the same value for market price.
- 6- and 12-month returns are based on total returns for the stock over trailing 6- and 12-month periods.

(z-score differences between decile 10 and decile 1)

GACP profile.[15] However, it seems that the stock market is sufficiently inefficient that it is possible to assemble a collective portfolio (like our decile 10) that indeed has the GACP profile.

Let us assess this evidence with some simple intuition. Look at the nature of the profile of decile 1 – risky, smaller capitalization, lower profitability and getting even worse, selling at relatively high prices compared to earnings, cash flow, sales and dividends, and with negative momentum over the past year. Compare this with the profile of decile 10 – lower risk, larger capitalization, higher profitability and getting even better, selling at low prices relative to earnings, cash flow, sales and dividends, with positive momentum over the past year. Ask yourself the following question. *Given a choice between investing in these two profiles, which would you choose?* We can safely say that the vast majority of investors would choose decile 10.[16] And, as it turns out, in the context of an *inefficient* market, this is the correct choice. Difficult as it may be to admit, the evidence suggests that this simple intuition is more powerful than any of the complex theories about expected return that can be found in the literature of modern finance!

There can be no question that risk goes down as you move from decile 1 through 10. In our view, the future "debate" will center on whether net return goes up or down as you move from 1 through 10. As we see in Table 23.2, in terms of raw return, it obviously goes up. Some may try to argue that the net of transactions costs does not go up *as much*.[17] The debate would then center on the level of risk aversion displayed here in the cross-section relative to the high level of risk aversion revealed in our longitudinal analysis as shown below.

In Fig. 23.1, we show the relationship between the daily changes in the implied volatility (the VIX, computed from options on the S&P 500 stock index) and the percentage changes in the index itself. Clearly, as the market's assessment of risk

[15] This raises an issue regarding the procedure by which money managers construct their portfolios. Stylized managers frequently sort stocks on the basis of some measure or measures of cheapness and then evaluate the sorted stocks on the basis of subjective considerations. Sorting procedures, whether applied to growth or value styles are limiting. To construct a GACP portfolio, we need to consider how each stock contributes to the profile of the final portfolio, much in the way a chef considers how each ingredient contributes to the taste of the final dish. Rather than sorting, portfolio managers might want to turn to linear programming to create attractive GACP opportunities.

[16] This assertion is based on an informal survey of many thousands of investors to whom Haugen has raised the issue in many speeches. Of course, until the issue was raised, the vast majority of these investors were never aware of the existence of GACP or its polar opposite DADP (decline at a dear price).

[17] Suppose that, in an efficient market, the only determinant of differences in cross-sectional expected return was market beta. As an investor you could invest in the market portfolio, which has a beta of one. This decision would likely require relatively low trading costs. Moving either to a higher or lower beta would require higher portfolio turnover to maintain the investment objective of higher or lower beta. The expected trading costs associated with increasing your personal utility by moving either to the left (lower beta) or to the right (higher beta) should be subtracted from both the expected returns to establish the true relationship between risk and return.

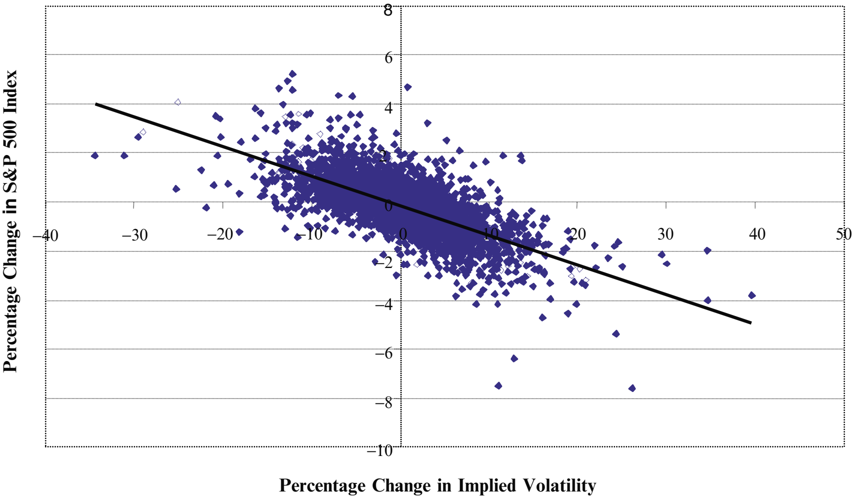


Fig. 23.1 Relationship between daily changes in implied volatility and *daily percentage changes in the S&P 500 (1/2/1990–6/13/2008)*. Daily percentage changes in the S&P 500 stock index are computed and plotted on the vertical scale. Percentage changes in the Chicago Board Options Exchange Volatility Index (VIX) are computed and plotted on the horizontal scale. The line is fitted using ordinary least squares

over the expiration period of the options goes up, the value of the index goes down.[18] A full 47% of the daily changes in the index can be explained by changes in the market's assessment of its volatility. As volatility goes up, investors require a higher rate of return on their stock investments. Given their expectations of future dividends, they can only get this by lowering the current market value of common stocks. Figure 23.1 reveals a high level of risk aversion on the part of investors. But where is the risk aversion in Table 23.3? Advocates of the efficient market hypothesis must reconcile Table 23.2, Table 23.3 and Fig. 23.1. In our opinion, this is impossible.

The combination of Table 23.2, Table 23.3, and Fig. 23.1 is a stake through the heart of the efficient market hypothesis.

Again, in order to accept the view of those who think the market is efficient, realized return, net of trading costs, must *fall dramatically* as we move from 1 to 10. Those who wish to maximize the utility by taking positions in the neighborhood of 10 must face significantly lower returns, net of trading costs, than those positioned in the neighborhood of decile 1. In our view this cannot be shown. It is not sufficient to show that a strategy whereby you "go short" decile 1 (and therefore, add trading costs to its returns) and "go long" decile 10 (subtracting trading costs from its returns) is non-profitable. *It must be shown that assuming higher (lower) risk garners higher (lower) returns, net of the trading costs associated with maintaining and managing the higher (lower) risk positions.*

[18] For those who wish to argue that the causation goes in the opposite direction from the return to implied volatility, we would ask, "Why would an extreme positive return cause a reduction in implied volatility while an extreme negative return causes an increase in implied volatility?"

We feel that our case against an efficient stock market is proved at this point. However, in the spirit of this volume, in Sect. 23.5 we shall see if these inefficiencies can be exploited after allowing for trading costs using a Markowitz-based investment strategy.

23.5 The Profitability of Managing Portfolios with the Expected Return Factor Model

Nearly 7 years after the publication of our original paper, Hanna and Ready (2005) wrote a paper in which they replicated, as closely as possible, our original results for the US markets. They then tested a trading strategy whereby deciles 1 and 10 are traded and the difference in returns is considered net of transactions costs. They contend the turnover associated with trading strategies using the expected return factor model eliminates its advantage relative to a strategy based on simple book-to-price or momentum.

In Sect. 23.8 of our original paper, we presented the results of an optimized (using a Markowitz-type procedure) trading strategy that was limited to the 1,000 largest US stocks. Portfolio turnover was limited to 20–40% per year and trading costs for these largest stocks were assumed to be a very generous 2% per round-trip. We showed that between 1979 and 1993, the difference in annualized return between a portfolio optimized to provide maximum return and the market index, net of transactions costs, was approximately 4%. This is what we argued to be the profitable predictability of the model. In Sect. 23.10, we provided a similar optimization analysis, net of transactions costs, globally. In this chapter, we expand our original optimized results to cover the extended time period and the five subperiods.[19]

In the optimizations, portfolio trading is controlled through a penalty function.

When available, the optimizations are based on the largest 1,000 stocks in the database. Estimates of the portfolio volatility are based on the full covariance matrix of the returns to the 1,000 stocks in the previous 24 months. Two years of monthly return data, from 1963 through 1964, is used to construct the initial portfolios. Estimates of expected returns to the 1,000 stocks are based on the factor model discussed above. The following constraints are applied to portfolio weights for each quarterly optimization:

(1) The maximum weight in a portfolio that can be assigned to a single stock is limited to 5%. The minimum is 0% (short selling is not permitted).
(2) The maximum invested in any one stock in the portfolio is three times the market capitalization weight or 0.25%, whichever is greater, subject to the 5% limit.

[19] We do not account for a 1-day trading lag in our analysis as did Hanna and Ready. This is because those who use the expected return factor model in practice re-estimate the model as of the close of trading at the end of the month and then rebalance their positions at the opening bell.

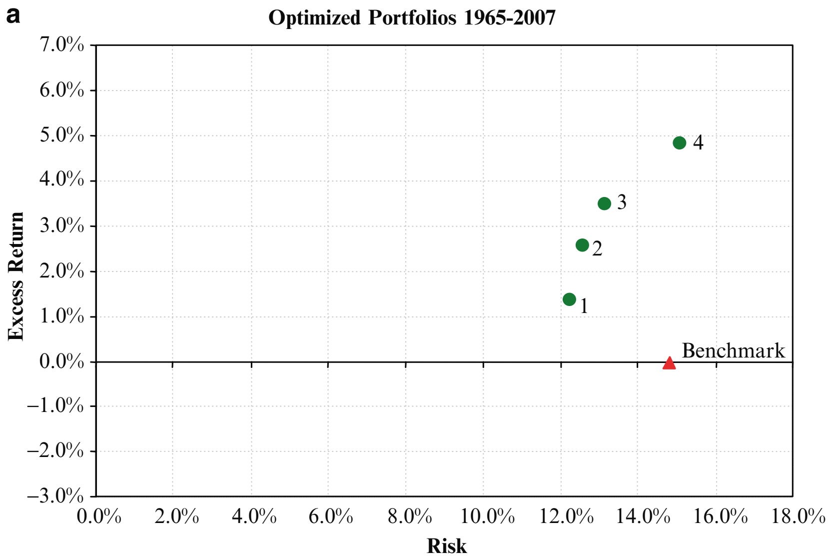


Fig. 23.2 Results of portfolio optimization. Four portfolios are optimized quarterly for the total period and (starting from scratch) for each of the subperiods. The subperiods are the same as in our other tests except for the fact that the 1963–1964 period is used to calculate the initial covariance matrix. The sample is restricted to the 1,000 largest stocks in our database. Estimates of portfolio risk are based on the full covariance matrix of historic returns over a 24-month window. Estimates of expected return are based on the ad hoc expected return factor model using information that was available at the beginning of each quarter. Monthly returns are linked for all portfolios

(3) The portfolio industry weight is restricted to be within 3% of the market capitalization weight of that industry (based on the two-digit SIC code.)
(4) Turnover in the portfolio is penalized through a linear cost applied to the trading of each stock. As a simplification, all stocks are subject to the same linear turnover cost although in practice portfolio managers use differential trading costs in their optimizations.

These constraints are designed to merely keep the portfolios diversified. Reasonable changes in the constraints do not materially affect the results.

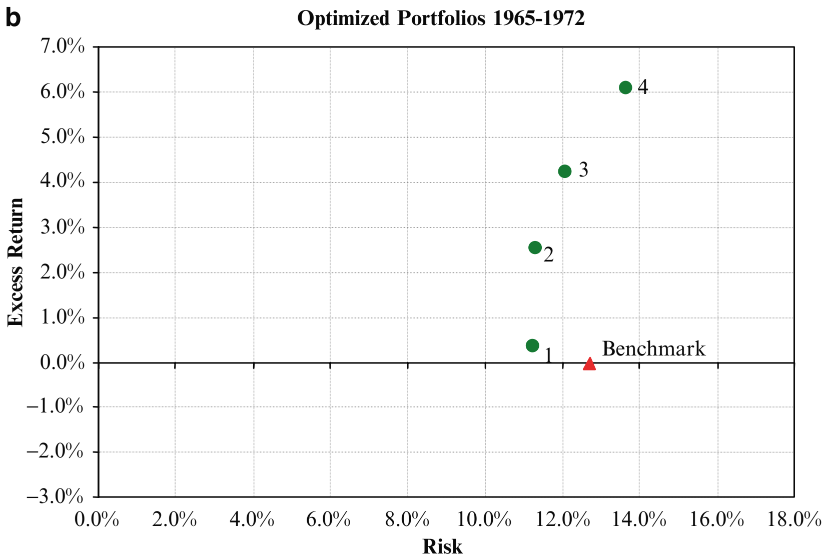
The portfolios are re-optimized quarterly.[20,21]

The performances of the four optimized portfolios across the total period[22] and the subperiods are presented in Figs. 23.2a–f and in Table 23.6. In the figures, the dots represent the four optimized portfolios. The triangle shows the position of the market benchmark. In the optimization process, we attempt to create a global minimum variance portfolio (which does not employ the expected return factor model at

[20] In our original paper, they also were re-optimized quarterly.

[21] With unconstrained optimization, with 24 monthly observations and 1,000 stocks, there is no unique solution. However, given the constraints provided above, unique solutions exist.

[22] We had to reserve 1963 and 1964 in order to calculate the covariance matrix for the first optimization for the first month (quarter) of 1965.



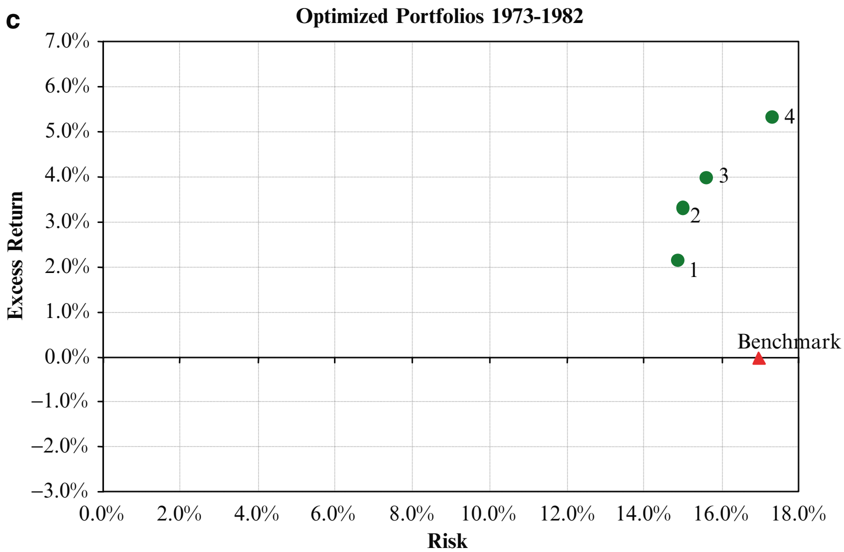
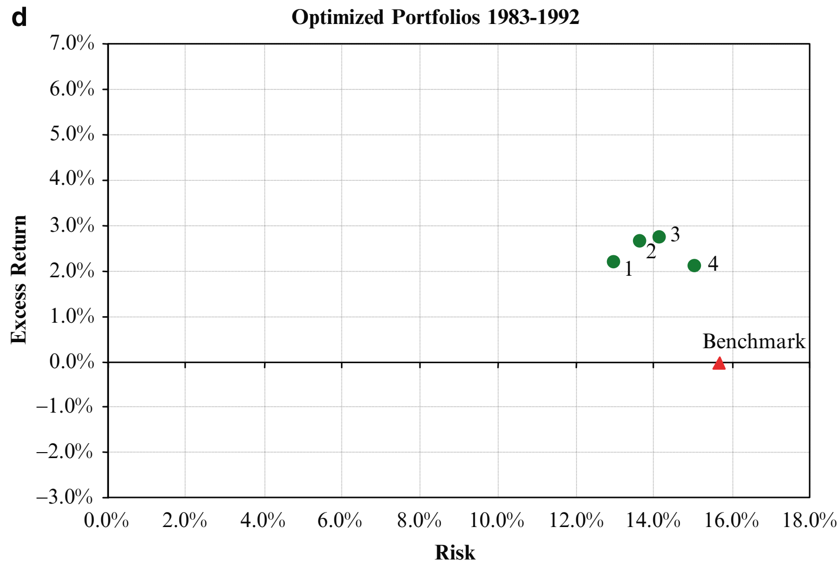


Fig. 23.2 (continued)

all) and three portfolios (that employs the model) that aim for successively higher expected return while minimizing the volatility.

Trading costs are not reflected in Fig. 23.2a–f. We leave it to the reader's judgment what the trading costs might be. However, in Table 23.5 we present the average annual turnover for each of the portfolios. To calculate what the round-



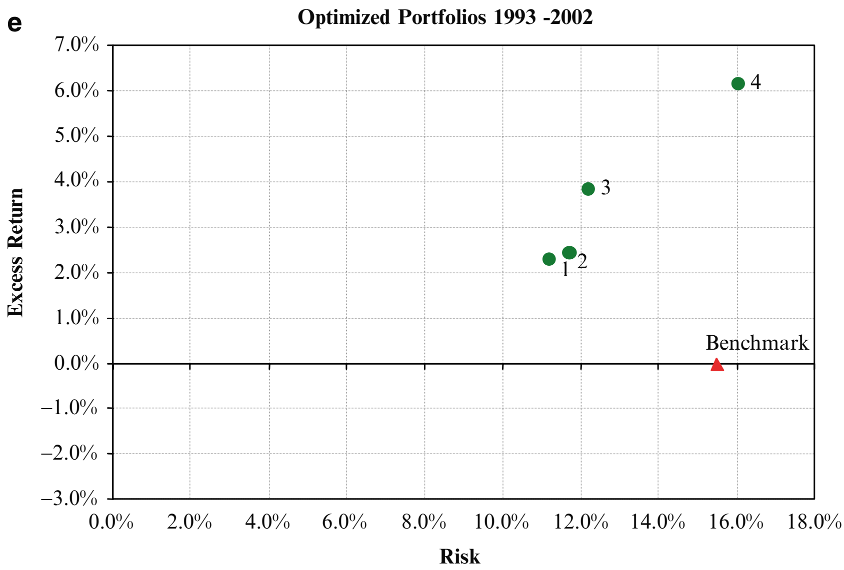


Fig. 23.2 (continued)

trip trading cost would be, in order to eliminate the spread between the optimized portfolios and the benchmark, simply divide the spread by the average annual turnover. Obviously, transactions costs would have to be unrealistically extreme to significantly close the gap between the high and low expected return portfolios.

Results for the optimizations in the subperiods are presented in Figs. 23.2b–f. Twenty-four months were reserved from the front-end of the first subperiod in order to calculate the initial covariance matrix.

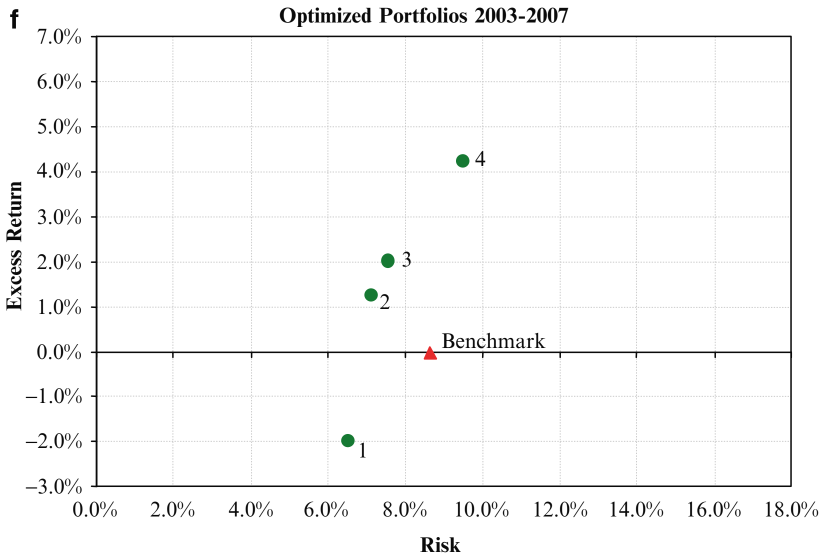


Fig. 23.2 (continued)

Table 23.5 Average annual turnover in the optimized portfolios

Period	1965–2007	1965–1972	1973–1982	1983–1992	1993–2002	2003–2007
Port. 1	11%	12%	13%	9%	12%	7%
Port. 2	38%	57%	38%	27%	41%	28%
Port. 3	62%	88%	61%	44%	64%	57%
Port. 4	80%	105%	84%	59%	80%	72%

Four portfolios are optimized quarterly for the total period and (starting from scratch) for each of the subperiods. The subperiods are the same as in our other tests except for the fact that the 1963–1964 period is used to calculate the initial covariance matrix. The sample is restricted to the 1,000 largest stocks in our database. Estimates of portfolio risk are based on the full covariance matrix of historic returns over a 24-month trailing window. Estimates of expected return are based on the ad hoc expected return factor model using information that was available at the beginning of each quarter. Annual turnovers, averaged over each year in the period, are provided in the cells of the table

Note the positions of the benchmark[23] relative to the global minimum variance portfolios. The positions reflect the fact that the payoff to risk was negative during the 45-year period. If we had constructed equally weighted portfolios of randomly selected stocks, and plotted their realized return against their volatility, our imagined scatter plot would have had a negative slope in the figures.

[23] The benchmark is the Russell 1,000 stock index for as long as it was in existence. Prior to this, the benchmark was the S&P 500 stock index.

Table 23.6 Expected returns and volatilities of the optimized portfolios and the market index

Portfolio	Statistics	1965–2007	1965–1972	1973–1982	1983–1992	1993–2002	2003–2007
Port. 1	Avg. excs. ret.	1.4%	.4%	2.1%	2.2%	2.3%	−2.0%
	Volatility	12.2%	11.2%	14.9%	13.0%	11.2%	6.5%
Port. 2	Avg. excs. ret.	2.6%	2.6%	3.3%	2.7%	2.4%	1.3%
	Volatility	12.6%	11.3%	15.0%	13.6%	11.7%	7.1%
Port. 3	Avg. excs. ret.	3.5%	4.3%	4.0%	2.7%	3.9%	2.0%
	Volatility	13.1%	12.1%	15.6%	14.1%	12.2%	7.6%
Port. 4	Avg. excs. ret.	4.8%	6.1%	5.3%	2.1%	6.2%	4.2%
	Volatility	15.1%	13.6%	17.3%	15.0%	16.1%	9.5%
Mkt. index	Avg. excs. ret	0	0	0	0	0	0
	Volatility	14.8%	12.7%	17.0%	15.7%	15.5%	8.6%

Four portfolios are optimized quarterly for the total period and (starting from scratch) for each of the subperiods. The subperiods are the same as in our other tests except for the fact that the 1963–1964 period is used to calculate the initial covariance matrix. The sample is restricted to the 1,000 largest stocks in our database. Estimates of portfolio risk are based on the full covariance matrix of historic returns over a 24-month trailing window. Estimates of expected return are based on the ad hoc expected return factor model using information that was available at the beginning of each quarter. Linked annualized average returns and annualized volatilities for the four optimized portfolios and for the market index are provided in the cells of the table

23.6 Summary

We find that the measures of current profitability and cheapness are overwhelmingly significant in determining the structure of the cross-section of stock returns. The statistical significance of risk is also overwhelming, but the payoff to risk has the wrong sign period after period. The riskiest stocks over measures including market beta, total return volatility, and residual volatility tend to have the lowest returns. We also find that the 1-year momentum pays off positively, and that last month's residual return and last quarter's total return pays off negatively. Strikingly, nearly all of the most significant factors over our total period are highly significant in our five subperiods, and all have the same signs as they did in the total period.

As in our earlier paper, the ad hoc expected return factor model is very powerful in predicting the future relative returns on stocks. High-return stock deciles tend to be relatively large companies with low risk and they have positive market price momentum. The profitability of high-return stocks is good and getting better. The low-return counterparts to these stocks have the opposite profile. A rational investor would likely find the high-return profile very attractive and the low-return profile very scary. Subsequently, they would tend to find their intuition about future return to have been proven correct.

In tests of rational trading strategies, where we account for trading costs, as reported in the original paper, the expected return factor model appears to be profitable net of trading costs. Profitability issues are irrelevant to our case against stock market efficiency. The case stands as proven.

Although attempts may be made, it is not likely that these results can be overturned without outrageous assumptions regarding the investor risk preferences, convoluted econometric techniques, or contrived, multifactor "risk adjustment procedures". The results presented here are the product of irrational behavior and the complexity and uniqueness of interaction on the part of investors. Like it or not, our results are out there for all to find and understand.

Given the evidence, and this evidence *will* be reproduced by others, the following conclusions are undeniable.

- The cross-sectional payoff to risk is highly negative
- The longitudinal payoff to risk is highly positive.
- The most attractive stock portfolios have the highest expected returns.
- The scariest stock portfolios have the lowest expected returns.

The stock market is inefficient. Case closed.

Acknowledgment We would like to thank Jan Bowler, Evan Einstein, Fred Elbert, and Tom Fees for their comments and research assistance.

References

Ang, Abdrew, Robert J. Hodrick, Yuhang Xing, and Xiaoyan Zhang, 2006, The Cross-section of Volatility and Expected Returns, *The Journal of Finance*, 259–299.

Fama, Eugene F. and Kenneth R. French, 1992, The Cross-section of Expected Stock Returns, *Journal of Finance*, 427–465.

Fama, Eugene F. and Kenneth R. French, 2008, Dissecting Anomalies, *Journal of Finance*, 1653–1678.

Fama, Eugene F. and James MacBeth, 1973, Risk, Return and Equilibrium: Empirical Tests, *Journal of Political Economy*, 607–636.

Hanna, J. Douglas and Mark J. Ready, 2005, Profitable Predictability in the Cross Section of Stock Returns, *Journal of Financial Economics*, 463–505.

Haugen, Robert A., *The Inefficient Stock Market – Second Edition* (Prentice Hall, Upper Saddle River, NJ, 2002).

Haugen, Robert A. *The New Finance – Overreaction, Complexity and Uniqueness, Third Edition* (Prentice Hall, Upper Saddle River, NJ, 2004).

Haugen, Robert A. and Nardin L. Baker, 1996, Commonality in the Determinants of Expected Stock Returns, *Journal of Financial Economics*, 401–439.

Haugen, Robert A. and A. James Heins, 1975, Risk and the Rate of Return on Financial Assets: Some Old Wine in New Bottles, *Journal of Financial and Quantitative Analysis*, 775–784.

Jegadeesh, Narasimhan and Sheridan Titman, 1993, Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency, *The Journal of Finance*, 65–92.

Lettau, Martin and Jessica A. Wachter, 2007, Why is the Long-horizon Equity Less Risky? A Duration-based Explanation of the Value Premium, *The Journal of Finance*, 55–92.

Chapter 24
Stock-Selection Modeling and Data Mining Corrections: Long-Only Versus 130/30 Models

John B. Guerard, Jr., Sundaram Chettiappan, and GanLin Xu

This study addresses several aspects of stock selection, portfolio construction, and data mining corrections and hypothesis testing of excess returns. Mean-variance, equally actively weighted, tracking-error-at-risk, and 130/30 portfolios are created and tests are conducted to find out whether the excess returns of these portfolios are statistically different from the average models that could have been used to build portfolios. Knowledge of earnings forecasts is an important input to the portfolio construction process. The portfolios constructed fulfill a global growth mandate, and the strategies work in EAFE plus Canada and the U.S. universes. The excess returns produced by the models are statically different from the average models used. The 130/30 strategy dominates the long-only strategy. Moreover, evidence is presented to show that these portfolios can be implemented and produce excess returns in the world of business.

Individual investors must be compensated for bearing risk. It seems intuitive that the risk of a security should be directly linked to its rate of return. Investors want to secure the maximum return for a given level of risk, or the minimum risk for a given level of return. The concept of risk–return analysis is known as the Efficient Frontier of Harry Markowitz (1952, 1959). If an investor can purchase a government security, which is backed by the taxing power of the Federal Government, then that government security is relatively risk free. The 90-day Treasury bill rate is used as the basic risk-free rate. Supposedly, the taxing power of the Federal Government eliminates the default risk of government debt issues, a liquidity premium is paid for longer term maturities because of the increasing level of interest-rate risk. Investors receive interest payments, as determined by the bond's coupon rate and may earn market price appreciation on longer term bonds if market rates fall, or sustain losses if market rates rise. From 1926 through 2003, Treasury bills returned 3.69%; longer term government bonds earned 5.28%; corporate bonds yielded 5.99%; and corporate stocks, as measured by the S&P 500 Index, earned 11.84% annually. Small stocks

J.B. Guerard, Jr. (✉) and S. Chettiappan
McKinley Capital Management, LLC, Anchorage, AK 99503, USA
e-mail: jguerard@mckinleycapital.com

G.Xu
Guided Choice, San Diego, CA, USA

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_24,
© Springer Science+Business Media, LLC 2010

averaged a 16.22% annual return over the corresponding period. The annualized standard deviations are 1.00%, 19.48%, and 29.66%, for Treasury bills, S&P 500 stocks, and small stocks respectively.

24.1 Constructing Efficient Portfolios

The Markowitz portfolio construction approach seeks to identify the efficient frontier, the point at which returns are maximized for a given level of risk, or minimize risk for a given level of return. The portfolio expected return, $E(R_{\mathrm{p}})$, is calculated by taking the sum of the security weights multiplied by their respective expected returns. The portfolio standard deviation is the sum of the weighted covariances.

$$E\left(R_{\mathrm{p}}\right) = \sum_{i=1}^{N} x_{\mathrm{i}} E\left(R_i\right) \tag{24.1}$$

$$\sigma_{\mathrm{p}}^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j \sigma_{ij} \tag{24.2}$$

The Markowitz framework measured risk as the portfolio standard deviation, its measure of dispersion, or total risk.[1] One sought to minimize covariances in the Markowitz framework, holding constant expected returns. However, as the number of securities, N, increased, the number of covariances increased faster, at covariances being N X N-1.

Elton and Gruber (2007) write a more modern version of the traditional Markowitz mean-variance problem as a maximization problem:

$$\theta = \frac{E(R_{\mathrm{p}}) - R_{\mathrm{F}}}{\sigma_{\mathrm{p}}} \tag{24.3}$$

[1] An examination of a three-asset portfolio construction problem follows.

$E(R_{\mathrm{p}}) = x_1\ E(R_1) + x_2\ E(R_2) + x_3\ E(R_3)$

let $x_3 = 1 - x_1 - x_2$

$E(R_{\mathrm{p}}) = x_1\ E(R_1) + x_2\ E(R_2) + (1 - x_1 - x_2)\ E(R_3)$

$$\begin{aligned}
\sigma_{\mathrm{p}}^2 &= x_1{}^2\sigma_1{}^2 + x_2{}^2\sigma_2{}^2 + \sigma_3{}^2 \mathrm{x}_3{}^2 + 2\mathrm{x}_1\mathrm{x}_2\sigma_{12} + 2\mathrm{x}_1\mathrm{x}_3\sigma_{23} + 2\mathrm{x}_2\mathrm{x}_3\sigma_{23} \\
&= x_1{}^2\sigma_1{}^2 + x_2{}^2\sigma_2{}^2 + (1 - x_1 - x_2)^2\ \sigma_3{}^2 + 2x_1x_2\ \sigma_{12} + 2x_1\ (1 - x_1 - x_2)\ \sigma_{13} \\
&\quad +2x_2\ (1 - x_1 - x_2)\ \sigma_{23} \\
&= x_1{}^2\sigma_1{}^2 + x_2{}^2\sigma_2{}^2 + (1 - x_1 - x_2)(1 - x_1 - x_2)\ \sigma_3{}^2 + 2x_1x_2\ \sigma_{12} + 2x_1\sigma_{13} - 2x_1{}^2\sigma_{13} \\
&\quad - 2x_1x_2\sigma_{13} + 2x_2\sigma_{23} - 2x_1x_2\sigma_{23} - 2x_2{}^2\sigma_{23} \\
&= x_1{}^2\sigma_1{}^2 + x_2{}^2\sigma_2{}^2 + \left(1 - 2x_1 - 2x_2 + 2x_1x_2 + x_1{}^2 + x_2{}^2\right)\ \sigma_3{}^2 + 2x_1x_2\ \sigma_{12} \\
&\quad +2x_1\sigma_{13} - 2x_1{}^2\sigma_{13} - 2x_1x_2\sigma_{13} + 2x_2\sigma_{23} - 2x_1x_2\sigma_{23} - 2x_2{}^2\sigma_{23}
\end{aligned}$$

$$\partial\sigma_{\mathrm{p}}^2 \Big/ \partial x_1 = 2x_1\left(\sigma_1{}^2 + \sigma_3{}^2 - 2\sigma_{13}\right) + x_2\ \left(2\sigma_3{}^2 + 2\sigma_{12} - 2\sigma_{13} - 2\sigma_{23}\right) - 2\sigma_3{}^2 + 2\sigma_{13} = 0$$

$$\partial\sigma_{\mathrm{p}}^2 \Big/ \partial x_2 = 2x_2\left(\sigma_2{}^2 + \sigma_3{}^2 - 2\sigma_{23}\right) + x_1\ \left(2\sigma_3{}^2 + 2\sigma_{12} - 2\sigma_{13} - 2\sigma_{23}\right) - 2\sigma_3{}^2 + 2\sigma_{23} = 0$$

where

$$E(R_{\mathrm{p}}) = \sum_{i=1}^{N} w_i = 1$$

$$\sigma_{\mathrm{p}} = \sum_{i=1}^{N} w_i^2 \sigma_i^2 + \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \sigma_{ij}$$

$$i \neq j$$

and R_{F} is the risk-free rate (90-day Treasury bill yield).

As in the initial Markowitz analysis, $\partial\theta/\partial w_i = 0$.

Thus, we may write the Elton-Gruber version of the Efficient Frontier as

$$\frac{\partial\theta}{\partial w_i} = -\left(\lambda w_1 \sigma_{1j} + \lambda w_2 \sigma_{2i} + \lambda w_3 \sigma_{3i} + \lambda w_i \sigma_i^2 + \lambda w_N \sigma_{Ni}\right) + \overline{R}_i - R_{\mathrm{F}} = 0. \tag{24.4}$$

Implicit in the development of the Capital Asset Pricing Model (CAPM) by Sharpe (1964), Lintner (1965), and Mossin (1966) is that investors are compensated for bearing not total risk, rather market risk, or systematic risk, as measured by the stock beta. The beta is the slope of the market model in which the stock return is regressed as a function of the market return. An investor is not compensated for bearing risk that may be diversified away from the portfolio.

The CAPM holds that the return to a security is a function of the security's beta.

$$R_{jt} = R_{\mathrm{F}} + \beta_j \left[E\left(R_{M_t}\right) - R_{\mathrm{F}}\right] + e_j, \tag{24.5}$$

where

R_{jt} = expected security return at time t;
$E(R_{M_t})$ = expected return on the market at time t;
R_{F} = risk-free rate;
β_j = security beta; and
e_j = randomly distributed error term.

An examination of the CAPM beta, its measure of systematic risk, from the Capital Market Line equilibrium condition follows:

$$\beta_j = \frac{\mathrm{Cov}(R_j, R_{\mathrm{M}})}{\mathrm{Var}(R_{\mathrm{M}})} \tag{24.6}$$

The Security Market Line (SML) shown in (24.4) is the linear relationship between return and systematic risk as measured by beta. The difficulty of measuring beta and its corresponding SML gave rise to extra-market measures of risk found in the work of King (1966), Farrell (1974), Rosenberg (1974, 1976, 1979), Stone (1974, 2002), Ross (1976), and Ross and Roll (1980). The BARRA risk model was developed in the series of studies by Rosenberg and thoroughly discussed in Grinhold and Kahn (1999).

The total excess return for a multiple-factor model (MFM) in the Rosenberg methodology for security j, at time t, dropping the subscript t for time, may be written as

$$E(R_j) = \sum_{k=1}^{K} \beta_{jk} \tilde{f}_k + \tilde{e}_j \tag{24.7}$$

The nonfactor, or asset-specific return on security j, is the residual risk of the security after removing the estimated impacts of the K factors. The term f is the rate of return on factor "k". A single-factor model, in which the market return is the only estimated factor, is obviously the basis for the CAPM. Accurate characterization of portfolio risk requires an accurate estimate of the covariance matrix of security returns. A relatively simple way to estimate this covariance matrix is to use the history of security returns to compute each variance, covariance, and security beta. The use of beta, the covariance of security and market index returns, is one method of estimating a reasonable cost of equity funds for firms.

However, the approximation obtained from simple models may not yield the best possible cost of equity. For example, the simple, single-index beta estimation approach suffers from the need to estimate a covariance matrix for the Russell 3000 stocks, requires a large amount of data, and is subject to estimation error. One would expect the estimated correlation between two stocks such as Microsoft and IBM to exceed the correlation between IBM and Dominion Resources because firms in similar sectors and/or industries tend to be more highly correlated than those in different industries. Furthermore, firms with similar characteristics, such as those in the same line of business, should have returns that behave similarly. For example, two technology firms with significant software lines of business should be more highly correlated than if one is a hardware firm and the other is a software provider. However, DuPont, IBM, and Dow Chemical will all have a common component in their returns, because they are all affected by news that affects the stock market, measured by their respective betas. The degree to which each of the three stocks responds to this stock market component depends on the sensitivity of each stock to the stock market component.

Specifically, DuPont (DD) and Dow Chemical are likely to respond to news affecting the chemical industry, whereas IBM and Dell probably will respond to news affecting the computer industry. The effects of such news may be captured by the average returns of stocks in the chemical industry and the petroleum industry. One can account for industry effects in the following representation for returns:

$$\begin{aligned} \tilde{r}_{\mathrm{DD}} = E[\tilde{r}_{\mathrm{DD}}] + \beta \cdot [\tilde{r}_{\mathrm{M}} - E[\tilde{r}_{\mathrm{M}}]] \\ +1 \cdot [\tilde{r}_{\mathrm{C}} - E[\tilde{r}_{\mathrm{C}}]] + 0 \cdot [\tilde{r}_{\mathrm{P}} - E[\tilde{r}_{\mathrm{DD}}]] + \mu_{\mathrm{P}}\text{'} \end{aligned} \tag{24.8}$$

where:

$\tilde{r}_{\mathrm{DD}}$ = DD's realized return,
$\tilde{r}_{\mathrm{M}}$ = the realized average stock market return,
$\tilde{r}_{\mathrm{C}}$ = the realized average return to chemical stocks,
$\tilde{r}_{P}$ = the realized average return to petroleum stocks,
E[.] = expectations,
β_{DD} = DD's sensitivity to stock market returns, and
μ_{DD} = the effect of DD specific news on DD returns.

This equation simply indicates that DD's realized return consists of an expected component and an unexpected component. The unexpected component depends on any unexpected events that affect stock returns in general $[\tilde{r}_{\mathrm{M}} - E[\tilde{r}_{\mathrm{M}}]]$, any unexpected events that affect the chemical industry $[\tilde{r}_{\mathrm{C}} - E[\tilde{r}_{\mathrm{C}}]]$, and any unexpected events that affect DD alone (μ_{DD}). Similar equations may be written for IBM and Dominion Resources.

The sources of variation in DuPont's stock returns thus are variations in stock returns in general, variations in chemical industry returns, and any variations that are specific to DuPont. Moreover, DuPont and Dow Chemical returns are likely to move together because both are exposed to stock market risk and chemical industry risk. DuPont, IBM, and Dominion Resources, on the other hand, are likely to move together to a lesser degree because the only common component in their returns is market return.

King (1966) began much of the multifactor modeling work with his analysis of industries. Rosenberg (1974) made substantial progress in estimating the covariance matrix of security returns to address the sources of co-movement in security returns. He identified the covariance matrix in terms of common sources in security returns, the variances of security specific returns, and estimates of the sensitivity of security returns to the common sources of variation in their returns, creating the BARRA risk model. Because the common sources of risk are likely to be much fewer than the number of securities, a much smaller covariance matrix is needed and a smaller history of returns is required. Moreover, because similar stocks will be more sensitive to similar common sources of risk, similar stocks will be more highly correlated than dissimilar stocks.

24.1.1 Alternative Multi-Beta Risk Models

Econometric estimation issues plagued early tests of the CAPM, see particularly Black, Jensen, Scholes (1972). Let us now examine several forms of multi-factor model returns. Bernell Stone (1974) developed a two-factor index model that molded equity returns as a function of an equity index and debt returns. More recently, Stone et al. (2002) developed an algorithm to generate portfolios that have similar stock betas (systematic risk), market capitalizations, divided yield, and sales growth cross-sections to enable one to access the excess returns of the analysts'

Table 24.1 Risk and return of mean-variance efficient portfolios, 1990–2001

Universe	Total active	t-Value	Asset selection	t-Value	Risk index	t-Value	Sectors	t-Value
RMC	1.98	1.37	0.99	0.86	0.97	1.45	−0.88	−0.97
R1000	2.47	2.52	1.85	2.12	0.82	2.13	−0.11	−0.23
R2500	7.76	4.37	6.48	3.96	1.61	2.85	−0.33	−0.62
R2000	9.68	5.83	8.81	5.57	0.90	2.36	−0.02	−0.07

RMC = Frank Russell Mid Cap Universe, R1000 = Frank Russell Largest 1000 Stock Universe, R2000 = Frank Russell Small Cap Universe, R2500 = Frank Russell Small and Mid Cap Universe
The CTEF variable produces statistically significant asset selection in the Russell 1000 universe during the 1990–2001 period

forecasts, forecast revisions, and breadth model as one moves from low (least preferred) to high (most preferred) securities with regard to a particular portfolio construction variable (i.e., CTEF). Excess returns similar to those shown in Table 24.1 can be produced with the Stone et al. (2002) algorithm during the 1982–1998 period.

Farrell (1974, 1997) estimated a four-factor, extra-market covariance model. He took an initial universe of 100 stocks (due to computer limitations) in 1973 and ran market models to estimate betas and residuals from the following market model:

$$R_{j_t} = a_j + b_j R_{M_t} + e_j \tag{24.9}$$

$$e_{j_t} = R_{j_t} - \hat{a}_j - \hat{b}_j R_{M_T} \tag{24.10}$$

The residuals of (24.9) should be independent variables. That is, after removing the market impact by estimating a beta, the residual of IBM should be independent of Dow Chemical, Merck, or Dominion Resources. The residuals should be, of course, independent in theory. Farrell (1974) examined the correlations among the security residuals of (24.10) and found that the residuals of IBM and Merck were highly correlated, but the residuals of IBM and Dominion Resources (then Virginia Electric & Power) were not.

Farrell used a statistical technique known as cluster analysis to create clusters, or groups, of securities having highly correlated market model residuals. He found four clusters of securities based on his extra-market covariance. The clusters contained securities with highly correlated residuals that were uncorrelated with residuals of securities in the other clusters. Farrell referred to his clusters as growth stocks (electronics, office equipment, drug, hospital supply firms, and firms with above-average earnings growth), cyclical stocks (metals, machinery, building supplies, general industrial firms, and other companies with above-average exposure to the business cycle), stable stocks (banks, utilities, retailers, and firms with below-average exposure to the business cycle), and energy stocks (coal, crude oil, and domestic and international oil firms). In 1976, Ross published his Arbitrage Theory of Capital Asset Pricing, which held that security returns were a function of four to five economic factors. In 1986, Chen et al. developed an estimated multifactor security return model based on the following:

$$\begin{aligned} R = a &+ b_{\mathrm{MP}}\mathrm{MP} + b_{\mathrm{DEI}}\mathrm{DEI} + b_{\mathrm{UI}}\mathrm{UI} + b_{\mathrm{UPR}}\mathrm{UPR} \\ &+ b_{\mathrm{UTS}}\mathrm{UTS}\,\mathrm{te}_t \end{aligned} \tag{24.11}$$

where

MP = monthly growth rate of industrial production,
DEI = change in expected inflation,
UI = unexpected inflation,
UPR = risk premium,
and UTS = term structure of interest rates.

Chen et al. defined unexpected inflation as the monthly first differences of the Consumer Price Index less the expected inflation rate. They used the "Baa and under"-rated bond return at the time less the long-term government bond return as the risk premium variable. The term structure variable was the long-term government bond return less the Treasury bill rates, known at time t-1 and applied to time t. When Chen, Ross, and Roll applied their five-factor model in conjunction with the value-weighted index betas during the 1958–1984 period, the index betas were not statistically significant, but the economic variables were.

The Stone, Farrell, and Chen, Ross, and Roll models used four to five factors to describe equity security risk. In addition, the models used different statistical approaches and economic models to control for risk. A simple question is now in order: If 4 or 5 betas produce better results than 2 or 3, why not estimate 20 or more betas? Importantly, the betas should be estimated on variables that are independent, or orthogonal, of the other variables. The estimation of 20 betas on orthogonal variables, 10 to 15 of which are not statistically significant, can produce expected returns for securities similar to four or five betas estimated on economic variables, or pre-specified variables.

24.1.2 Security Weights and Portfolio Construction

The estimation of security weights in a portfolio is the primary calculation of Markowitz's portfolio management approach. The issue of security weights is now considered from a different perspective. As previously discussed, the security weight is the proportion of the portfolio's market value invested in the individual security.

$$w_{\mathrm{s}} = \frac{\mathrm{MV_s}}{\mathrm{MV_p}} \tag{24.12}$$

where

w_{s} = portfolio weight insecurity s,
$\mathrm{MV_s}$ = value of security s within the portfolio,
and $\mathrm{MV_p}$ = the total market value of portfolio.

The active weight of the security is calculated by subtracting the security weight in the (index) benchmark, b, from the security weight in the portfolio, p.

$$w_{\mathrm{s,p}} - w_{\mathrm{s,b}} \tag{24.13}$$

That is, if IBM has a 3% weight in the portfolio while its weight in the benchmark index is 2.5%, then IBM has a positive, 0.05% active weight in the portfolio. The portfolio manager has an active, positive opinion of securities on which he or she has a positive active weight and a negative opinion of those securities with negative active weights.

Markowitz analysis (1952, 1959) and its efficient frontier minimized risk for a given level of return. Risk can be measured by a stock's volatility, or the standard deviation in the portfolio return over a forecast horizon, normally 1 year.

$$\sigma_{\mathrm{p}} = \sqrt{E(r_{\mathrm{p}} - E(r_{\mathrm{p}}))^2} \tag{24.14}$$

Blin and Bender created an APT, Advanced Portfolio Technologies, Analytics Guide (2005), which built upon the mathematical foundations of their APT system, published in Blin et al. (1997). The following analysis draws upon the APT analytics. Volatility can be broken down into systematic and specific risk:

$$\sigma_{\mathrm{p}}^2 = \sigma_{\beta\mathrm{p}}^2 + \sigma_{\varepsilon\mathrm{p}}^2 \tag{24.15}$$

where

σ_{p} = total portfolio volatility,
$\sigma_{\beta_{\mathrm{p}}}$ = systematic portfolio volatility,
and $\sigma_{\varepsilon\mathrm{p}}$ = specific portfolio volatility.

Blin and Bender created a multifactor risk model within their APT risk model based on forecast volatility.

$$\sigma_{\mathrm{p}} = \sqrt{52\left(\sum_{c=1}^{c}\left(\sum_{i=1}^{s} w_i \beta_{i,c}\right)^2 + \sum_{i=1}^{s} w_i^2 \varepsilon_{i,w}^2\right)} \tag{24.16}$$

where

σ_{p} = Forecast Volatility of Annual Portfolio Return,
C = Number of Statistical Components in the Risk Model,
w_i = Portfolio weight in security i,
$\beta_{i,c}$ = The loading (beta) of security i on risk component c,
$\varepsilon_{i,w}$ = Weekly Specific Volatility of Security i.

The Blin and Bender (1995) systematic volatility is a forecast of the annual portfolio standard deviation expressed as a function of each security's systematic APT components.

$$\sigma_{\beta\mathrm{p}} = \sqrt{52\sum_{c=1}^{c}\left(\sum_{i=1}^{s} w_i \beta_{i,c}\right)^2} \tag{24.17}$$

Portfolio-specific volatility is a forecast of the annualized standard deviation associated with each security's specific return.

$$\sigma_{\varepsilon \mathrm{p}} = \sqrt{52 \sum_{i=1}^{s} w_i^2 \varepsilon_{i,w}^2} \tag{24.18}$$

Tracking error, te, is a measure of volatility applied to the active return of funds (or portfolio strategies) that are indexed against a benchmark, which is often an index. Portfolio tracking error is defined as the standard deviation of the portfolio return less the benchmark return over 1 year.

$$\sigma_{\mathrm{te}} = \sqrt{E\left[\left((r_{\mathrm{p}} - r_{\mathrm{b}}) - E(r_{\mathrm{p}} - r_{\mathrm{b}})\right)^2\right]} \tag{24.19}$$

where

σ_{te} = annualized tracking error,
r_{p} = actual portfolio annual return,
r_{b} = actual benchmark annual return.

The APT-reported tracking error is the forecast tracking error for the current portfolio versus the current benchmark for the coming year. The reader is referred to the APT analysis presented in Chapter 2.

Blin and Bender (1997–1997) estimated a 20-factor beta model of covariances based on 2.5 years of weekly stock returns data. The Blin and Bender Arbitrage Pricing Theory (APT) model followed the Ross factor modeling theory, but Blin and Bender estimated betas from at least 20 orthogonal factors. Blin and Bender never sought to identify their factors with economic variables. The APT 20–24 beta modeling process is an alternative to the multi-factor models of Cohen et al. (1967), Elton et al. (1970), Ross and Roll (1980), Chen et al. (1986), Dhrymes et al. (1984), Conner et al. (1988) and 1995, and Guerard et al. (1997).

24.2 Variable Selection, Testing, and Portfolio Construction

There are a seemingly infinite number of investment strategies. One of the oldest, best known, and most practiced strategies is value investing, in which an investor or manager seeks to identify undervalued assets. The concept of value investing is generally attributed to Graham and Dodd (1934). They criticized common stock investing that ignored stock price relative to measures of fundamental value and established many concepts of fundamental value-focused investment analysis. In advocating a low price/earnings P/E multiple strategy, Graham and Dodd noted that, prior to the 1927–1929 bull market, a reasonable PE multiple was 10. In the 1927–1929 bull market, many stocks had PE multiples of 25 to 40, considerably exceeding

Graham and Dodd's prudent maximum of 16. While the Graham-Dodd analysis involved many qualitative and quantitative considerations (such as avoiding equities of firms with excessive debt, a low stock price, poor liquidity, and those without positive earnings track records), they focused much of their stock-level value assessment on three value attributes: dividends, earnings, and measures of net asset value.

Graham and Dodd advocated a history of positive and growing dividends. However, they focused their valuation on future earnings. Although a history of positive and growing earnings was desirable, past earnings were relevant to value only to the extent that they predicted future earnings. Graham and Dodd were reluctant to forecast earnings more than 4 years into the future. The crux of their earnings-focused value assessment was the P/E ratio. All else being equal, a lower P/E ratio implied a better investment value. There are two dimensions to the all-else-being-equal concept: comparison with peer companies and comparison with a company's own past P/E ratios[2].

In addition to the extensive literature on the impact of individual value ratios on the cross section of stock returns, even more literature exists on the use of value ratios to predict stock returns (or relative returns), risks, and special situations such as merger candidates or financial misreporting. This review, however, is restricted to a subset of general stock return forecasting, namely one-step[3] regression-estimated composites of three or more value ratios, especially those that go beyond using

[2] The influence of Graham and Dodd, primarily through the low PE multiple, continued. Students in the 1970s were studying their conclusions in a standard security analysis, investments, and portfolio management text, *Security Analysis and Portfolio Management*, by Latane et al. (1975). The low PE approach is often referred to as an area of "fundamental analysis." An investor uses income statement and balance sheet data to estimate an intrinsic value. Such data include sales, earnings, book value, and risk characteristics. Haugen (1996, 1999, 2001) continues the treatment of the Graham and Dodd approach, examining 12 of the most important factors in the U.S. equity markets and in Germany, France, Great Britain, and Japan. The book-to-price, earnings-to-price, sales-to-price, and cash flow-to-price variables were among the highest mean payoff variables in the respective countries. Moreover, Graham and Dodd became the standard reference of the value-investing practitioner community due, in large part, to A. Bernhard (1959), the founder of *Value Line*, and his former senior editor, David Dremen, who rose to become chairman of Dremen Value Management, LLC. Dremen (1979, 1998) is the ultimate contrarian. Dremen (1988) showed the vast outperformance of the low P/E, low P/CF, low P/BV (book value), low P/D (dividends) of the 1500 largest stocks in the U.S., 1970 –1996 (pages 154–155). These contrarian variables averaged almost 350 basis points, annually, over the market return, though Dremen did not account for the riskiness of the strategies.

[3] Rather than the one-step direct forecast of stock returns considered here, most of the literature (including the framework presented in investment management texts) and a majority of the stock-return forecasting using valuation multiples is in the context of a two-step return forecast in which an analyst predicts both a future value of a variable such as earnings and an associated future value multiple for that variable such as a future price-earnings ratio. The most common version is a prediction of future earnings and of a future earnings multiple. These two predictions imply a prediction future value. Under the assumption that the current price will converge toward this predicted future value, there is an implied prediction of a gain return. Given a prediction of future dividends, there is an implied stock return forecast. In the conventional two-step procedure, this research uses a direct one-step prediction of return as a composite of current and relative value ratios.

just one or two of the standard value ratios [dividend-to-price (DP), earnings-to-price (EP), and/or book-to-price (BP)] to include as well the cash-to-price ratio (CP) and/or the sales-to-price ratio (SP). The major papers in this area include Chan et al. (1991), Bloch et al. (1993); Lakonishok et al. (1994); and Guerard et al. (1997).

Bloch et al. (1993) used the estimated regression coefficients to construct a rolling horizon return forecast. These predicted returns and predictions of risk parameters were then input to a mean-variance optimizer [see Markowitz (1987)] to create mean-variance efficient portfolios in both Japan (first section, non-financial Tokyo Stock Exchange common stocks, January 1975 to December 1990) and the United States (the 1000 largest market-capitalized common stocks, November 1975 to December 1990).[4] Bloch et al. (1993) found that Markowitz (1987) mean-variance efficient portfolios using the lower EP values in Japan underperformed the universe benchmark, whereas BP, CP, and SP (sales-to-price, or sales yield) variables outperformed the universe benchmark. The optimized portfolios using BP, CP, SP, and EP variables outperformed the U.S. S&P 500 Index, giving support to the Graham-Dodd concept of using the relative ranking of value-focused fundamental ratios to select stocks.

Bloch et al. (1993) used relative ratios as well as current ratio values. Not only might an investor desire to purchase a low P/E stock, but one might desire to purchase a low P/E stock when the P/E is at a relatively low value compared to its historical value, in this case a low relative to its average over the last 5 years. Bloch et al. (1993) reported several results.

Guerard et al. (1997) studied the intersection of Compustat, CRSP, and I/B/E/S data bases. This study built on the fundamental forecasting work in Bloch et al. (1993) in two ways: it added to the Bloch et al. eight-variable regression equation a growth measure, then added three measures of analyst forecasts and forecast revisions from the I/B/E/S data base, namely consensus analysts' forecasts, forecast revisions, and the direction (net up or down) of the forecast revisions. Adding I/B/E/S variables to the eight value ratios produced more than 2.5% of additional annualized return. The finding of significant predictive performance value for the three I/B/E/S variables indicates that analyst forecast information has value beyond purely statistical extrapolation of past value and growth measures. Possible reasons for the additional performance benefit could be that analysts' forecasts and forecast revisions reflect information in other return-pertinent variables, discontinuities from past data, and/or serve as a quality screen on otherwise out-of-favor stocks. The quality screen idea is a confirmation of the Graham and Dodd argument that value ratios should be used in the context of the many qualitative and quantitative factors that they argue are essential to informed investing.

[4] The use of non-financial stocks led to a customized index for the Markowitz Global Portfolio Research Group (GPRD) analysis. The Chan et al. and an initial Guerard presentation occurred in September 1991 at the Berkeley Program in Finance, Santa Barbara, on Fundamental Analysis. Bill Ziemba presented a very interesting study comparing U.S. and Japanese fundamental strategies at the same Berkeley Program meeting.

In terms of relative predictive value, Guerard et al. (1997) found the EP, CP, SP, and RSP variables to be more important than the BP variable. Thus, a large amount of testing both value and growth variables exists in the non-U.S. and U.S. markets. For example, one might estimate information coefficients, the correlation coefficient between the current strategy score, and subsequent monthly (or quarterly) total returns. Approximately 10,000 securities are covered by analysts in the I/B/E/S database in the current non-U.S. universe. An investment manager probably cannot invest in all of these securities with equal effectiveness. A liquidity screen identified a universe of approximately 2700 securities during much of the October 1995 through April 2008 period that was an effective universe for creating efficient portfolios that could be executed.[5]

The ICs of the variables in our non-U.S. analysis for all securities and liquidity-screened securities is shown in Table 24.2. Most of the variables are statistically significant; however, a variable cannot be effectively used in portfolio construction if the returns are dependent upon systematic sources of returns. MQ, a variable combining price momentum and analysts' revisions, is particularly strong in U.S. and non-U.S. markets, produces the highest ICs in the all security EAFE plus Canada universe, and produces a statistically significant IC in the liquidity-screened universe. The price momentum (PM) component of the MQ variable is relatively

Table 24.2 Information coefficients (EAFE + Canada universe 10/1995–4/2008)

	Universe		Universe	
		All		Liquidity-screened
Variable	IC	t(IC)	IC	t(IC)
MQ	0.06	39.8	0.04	13.6
EP	0.05	44.1	0.07	29.2
PM	0.05	42.2	0.03	13.3
CP	0.05	39.5	0.05	24.0
FEP2	0.04	32.0	0.04	18.0
FEP1	0.03	28.1	0.03	16.3
RV2	0.02	20.1	0.02	9.3
RV1	0.02	16.7	0.01	7.1
BP	0.02	12.8	0.03	11.1
SP	0.00	2.6	0.01	6.0
FGRI1	0.00	−0.2	−0.01	−2.0
FGRI2	−0.01	−3.6	−0.02	−4.2
LTG	−0.02	−6.1	−0.01	−3.9

IC = information coefficient, *t(IC)* = *t* statistic of the information coefficient, *MQ* = McKinley quant score, *EP* = earnings/price, *PM* = price momentum, *CP* = cash flow/price, *FEP1* = forecasted FY1 EPS/price, *FEP2* = forecasted FY2 EPS/price, *RV1* = revisions to FY1 EPS, *RV2* = revisions to FY2 EPS, *BP* = book value/price, *SP* = sales/price, *LTG* = projected long-term EPS forecast

[5] Please see McKinley Capital Management, Inc., 2006. "An Assessment of the McKinley Capital Non-U.S. Developed Growth and Non-U.S. Growth Portfolios," November.

Table 24.3 Data mining corrections models (EAFE + Canada long-only strategy liquidity-screened universe 10/1995–4/2008)

Variable	GM	STD	ShR	TError	IR
FEP1	17.76	15.79	0.88	6.03	1.19
CP	14.95	17.43	0.64	7.57	0.59
FEP2	14.82	17.32	0.63	8.20	0.49
BP	12.94	21.17	0.43	12.30	0.16
MQ	12.47	17.84	0.48	8.19	0.56
PM	11.77	17.15	0.46	7.92	0.48
CTEF	10.74	15.01	0.46	5.23	0.13
LTG	10.01	15.80	0.39	5.41	0.14
SP	9.97	20.95	0.29	11.47	−0.11
RV1	9.94	16.51	0.37	5.48	0.12
RV2	8.97	15.58	0.33	4.85	−0.21
FGRI1	8.55	16.41	0.28	6.82	−0.14
Benchmark	8.12	14.26	0.30		
FGRI2	6.04	17.16	0.13	7.00	−0.40

GM = geometric mean, *STD* = annualized standard deviation, *ShR* = Sharpe ratio, *TError* = tracking error, IR = information ratio, *MQ* = McKinley quant score, *PM* = price momentum, *BP* = book value/price, CP = cash flow/price, *EP* = earnings/price, *SP* = sales/price, *FEP1* = forecasted FY1 EPS/price, *FEP2* = forecasted FY2 EPS/price, *RV1* = revisions to FY1 EPS, *RV2* = revisions to FY2 EPS, *CTEF* = average (revisions, forecasted EPS yields, breadth), *LTG* = projected long-term EPS forecast, *FGRI1* = FY1 EPS forecast/last year's reported EPS, *FGRI2* = FY2 EPS forecast/last year's reported EPS

affected by liquidity concerns. Analysts revisions are an important determinant of stock returns, see Brown (2008) and Ramnath et al. (2008).

Table 24.3 shows optimized portfolios in the non-U.S. universe for October 1995 through April 2008; Table 24.4 shows optimized portfolios in the Russell 3000 U.S. securities for January 1997 through April 2008. Statistically significant support is found for the price momentum and analysts' revisions variables in U.S. and non-U.S. equity universes. MQ, a variable combining price momentum and analysts' revisions, is particularly strong in U.S. and non-U.S. markets, producing geometric means (GM), Sharpe Ratios (ShR), and information ratios (IR) that significantly exceed those of the benchmark. In portfolio management, one seeks to maximize the Geometric mean of portfolios [Markowitz (1976)]. Similar evidence supports book value, cash flow, and sales variables in U.S. and non-U.S. markets.

Not all variables are statistically significant and useful in model creation. For example, the 1- and 2-year-ahead analysts' forecasted growth variables of Elton et al. (1981) offer near or below benchmark GMs, ShRs, and IRs.

The MQ variable can be used to effectively select models and create portfolios in non-U.S. and U.S. universes. Equally actively weighted (EAW) portfolios have been created using the Markowitz enhanced index tracking (EIT) methodology. A 2% upper limit is used in this analysis. Support for the estimation of efficient EIT portfolios can be found in Guerard, Takano, and Yamane (1993).

Table 24.4 Data mining corrections models (U.S. Russell 3000 strategy 1/1997–4/2008)

Variable	GM	STD	ShR	TError	IR
MQ	16.15	19.06	0.66	7.96	0.96
PM	13.97	17.31	0.60	8.02	0.83
SP	12.91	18.39	0.39	14.53	0.43
RV1	10.91	18.17	0.40	6.01	0.67
CTEF	10.61	17.10	0.41	6.73	0.59
CP	10.06	22.43	0.29	12.60	0.30
RV2	9.90	18.39	0.34	5.62	0.52
FEP1	7.41	22.46	0.17	10.49	0.11
Benchmark	5.49	14.96	0.30		
BP	5.28	23.97	0.07	14.15	−0.01
FEP2	1.65	24.13	−0.08	12.40	−0.22
LTG	0.30	30.94	−0.11	13.97	0.14
FGRI1	8.55	16.41	0.28	6.82	−0.14
FGRI2	6.04	17.16	0.13	7.00	−0.40

GM = geometric mean, *STD* = annualized standard deviation, *ShR* = Sharpe ratio, *TError* = tracking error, IR = information ratio, *MQ* = McKinley quant score, *PM* = price momentum, *BP* = book value/price, CP = cash flow/price, *EP* = earnings/price, *SP* = sales/price, *FEP1* = forecasted FY1 EPS/price, *FEP2* = forecasted FY2 EPS/price, *RV1* = revisions to FY1 EPS, *RV2* = revisions to FY2 EPS, *CTEF* = average (revisions, forecasted EPS yields, breadth), *LTG* = projected long-term EPS forecast, *FGRI1* = FY1 EPS forecast/last year's reported EPS, *FGRI2* = FY2 EPS forecast/last year's reported EPS

A tradeoff curve is estimated for the liquidity-screened EAFE plus Canada universe for the October 1995 through April 2008 period. There is virtually no difference in the TaR-estimated tradeoff curve using (24.33) and the corresponding EAW tradeoff curve using a variation of the Markowitz (2000) EIT methodology in which a weight cannot deviate by more than 2% from the universe benchmark (see Chart 24.1).

Investment managers may use a risk-control process that includes size and sector constraints. They may seek to dampen volatility by imposing a constraint, for example, that the portfolio's size will never fall below 50% of the universe benchmark (as measured by market capitalization) or deviate by no more 5% in sectors.

Table 24.5 shows the results of imposing such size and sector constraints on many of the non-U.S. growth factors examined in Table 24.2 using the January 1997 through December 2007 period. The simulated portfolio was further constrained, using the APT Associates risk model, by 300 basis points of transaction costs (20 basis points of hard commission and 280 basis points of market impact), and approximately 100% annual turnover was assumed (for the long and short portfolios separately) to approximate an actual investment process.

The MQ variable is effective in the size and sector constrained EAFE plus Canada universe in the January 1997 through December 2007, a result consistent with the actual experience of a quantitative manager.

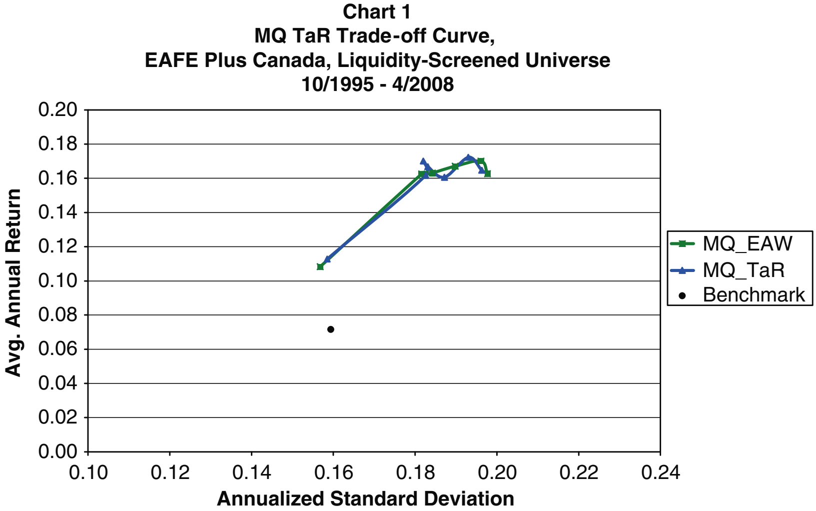


Chart 24.1 MQ TaR trade-off curve, EAFE plus Canada, liquidity-screened universe 10/1995–4/2008

Table 24.5 Size and sector constrained optimized portfolios (EAFE + Canada Universe 1/1997–12/2007)

Variable	GM	STD	ExR	ShR	TError	IR
MQ	11.17	17.30	2.92	0.43	6.16	0.47
FEP1	9.43	14.52	−0.95	0.39	5.40	−0.18
FEP2	9.00	15.14	−1.65	0.35	5.62	−0.29
BP	8.79	15.81	−1.06	0.32	5.64	−0.19
PM	8.50	16.83	0.25	0.28	7.09	0.03
EP	8.01	14.56	−1.75	0.29	4.79	−0.36
RV1	7.14	16.49	−1.94	0.20	5.01	−0.39
CP	7.11	16.84	−2.92	0.20	5.76	−0.51
LTG	6.86	17.60	−1.38	0.18	5.42	−0.25
Benchmark	6.78	15.52		0.20		
SP	6.52	16.77	−3.42	0.16	5.38	−0.64

GM = geometric mean, STD = annualized standard deviation, *ExR* = excess returns versus EAFE growth benchmark, *ShR* = Sharpe ratio, *TError* = tracking error, *IR* = information ratio, *MQ* = McKinley quant score, *FEP1* = forecasted FY1 EPS/price, *FEP2* = forecasted FY2 EPS/price, *BP* = book value/price, *PM* = price momentum, *EP* = earnings/price, *RV1* = revisions to FY1 EPS, *CP* = cash flow/price, *LTG* = projected long-term EPS forecast, *SP* = sales/price

24.3 Data Mining Corrections: Did You Get Lucky?

In the (practical) world of Wall Street, it is conventional wisdom to cut your historical backtested excess returns in half; that is, if your backtested excess returns (the portfolio geometric mean returns less the geometric mean of the benchmark) was 6%, or 600 basis points, an investment manager and/or a client might expect

3% excess returns in the future. In January 1991, Harry Markowitz and his Daiwa Global Portfolio Research Department launched Fund Academy, a Japanese-only Tokyo-based investment strategy. In its first year, ending December 1991, the Daiwa Portfolio Optimization System (DPOS) out performed the benchmark by some 700 basis points. Markowitz asked the question "Did we get lucky?" The obvious answer is "no, we imported a U.S.-estimated strategy run successfully at Drexel Burnham Lambert to Japan, its backtested excess returns were 50% higher in Japan than in the U.S., and its 700 basis points of excess returns were not that less than its 1000 basis points of backtested out-performance during the 1974–1990 period." That answer, while true, was not scientific. Ganlin Xu, the Daiwa Global Research Department mathematician developed a testing methodology that Markowitz and Xu (1994) published in the *Journal of Portfolio Management*. Let us trace the development of the Markowitz and Xu model and estimate the Data Mining Corrections estimator for the long-only and 130/30 strategies. In Bloch et al. (1993), some 200 historic U.S. and Japanese equity model simulations were reported. The best backtested DPOS strategy was dominant in U.S. and Japanese markets. The DPOS Japanese-only strategy was funded and the level of expected future out-performance was a question (always) asked by clients.

Let $\mathrm{GM_b}$ be the backtested geometric best of the "best" historic simulation during T periods. Markowitz and Xu (1994) work with the logarithm of the geometric mean.

$$g_{\mathrm{b}} = \log_{\mathrm{e}}(1 + \mathrm{GM_b})\,. \tag{24.20}$$

The Markowitz and Xu Data Mining Corrections (DMC) test assumes that the T period historic returns were identically and independently distributed (i.i.d.), and it is assumed that future returns are drawn from the same population (also i.i.d.). Since we test many models, not just on g_{b} is no longer the best unbiased estimate of T the true, underlying population g_{b}.

Markowitz and Xu (1994) set y_{it} as the logarithm of one plus the return for the ith portfolio selection model in period t. y_{it} is of the form

$$y_{it} = \mu_i + z_t + \varepsilon_{it}, \tag{24.21}$$

where

μ_i is a model effect,
z_t is a period effect,
and ε_{it} is a random deviation.

In Markowitz and Xu (1994) Model I, it is assumed that z_t is observable and assumed to be the return of a market index. In this case, r_{it} is an excess return of model i:

$$r_{it} = y_{it} - z_{it} = \mu_i + \varepsilon_{it}. \tag{24.22}$$

The random deviation ε_{it}, has a zero mean and is uncorrelated with μ_i and other ε_{js}:

$$\mathrm{E}\left(\varepsilon_{it}\right) = 0 \tag{24.23}$$

$$\operatorname{cov}\left(\mu_i, \varepsilon_{jt}\right) = 0 \quad \text{for all } j \text{ and } t \tag{24.24}$$

$$\operatorname{cov}\left(\varepsilon_{it}, \varepsilon_{js}\right) = 0 \quad \text{for all i} \neq \text{j or s} \neq \text{t.} \tag{24.25}$$

Model I excess returns are assumed to be independent of time and other error terms of other models.

Markowitz and Xu (1994) posed a Model II in which z_t, the market index return, is not observable. They did not estimate Model II in their original work, and we will not estimate it in this analysis. We proceed to estimate Model III. In Model III,

$$y_{it} = \mu_i + \varepsilon_{it}, \quad \text{as in.} \tag{24.26}$$

Markowitz and Xu did not require that model to be independent of one another. Thus, in Model III, cov $(\varepsilon_{it}, \varepsilon_{jt})$ need not be zero. Thus, Model III is not only the general case (Model I being a special case of Model III), but Model III is also consistent with testing in business-world portfolio construction and testing. Finally, the appropriate estimate of μ_{it} in Model I is not the average return

$$\bar{r}_i = \frac{\sum_{t=1}^{\mathrm{T}} y_{it}}{\mathrm{T}}, \tag{24.27}$$

$$\text{but rather} \quad \bar{r} = \sum_{i=1}^{\mathrm{T}} r_i / n. \tag{24.28}$$

The estimate of μ_i is regressed back to the average return (the grand average),

$$\hat{\mu} = \bar{r} + \beta\,(\bar{r}_i - \bar{r}), \tag{24.29}$$

where $0 < \beta < 1$.

The best linear estimate of the unknown μ_i is

$$\hat{\mu}_i = E\,\mu + \beta\,(\bar{r}_i - E\mu) \tag{24.30}$$

$$\beta = \frac{\operatorname{cov}(\bar{r}_i, \mu)}{\operatorname{Var}(\bar{r}_i)}. \tag{24.31}$$

Thus, β is the regression coefficient of μ_i as a function of r_i. Does the return of the best model deviate (statistically significant) from the average model return? The best linear unbiased estimate of the expected model return vector, μ is:

$$\hat{\mu} = E(\mu)e + \operatorname{Var}(\mu)\left[\tfrac{1}{T}C + \operatorname{Var}(\mu)I\right]^{-1} \times (\tilde{y} - E(\mu)e) \tag{24.32}$$

$$C = \operatorname{cov}\left(\varepsilon_e, \varepsilon_j\right)$$

The Markowitz-Xu DMC test did not use a "holdout period," as they can be routinely data mined as well. That is, one can vary the estimation and holdout periods to generate the desired conclusion. In Markowitz and Xu (1994), they tested the DPOS strategies in Bloch et al. (1993), and the best model is illustrated in Chart 24.1. Markowitz and Xu reported a Model III β of 0.59, which was statistically significant. In the King's English, approximately 59% of the excess returns could be expected to continue.

The MQ variable passes the Data Mining Corrections test criteria for both U.S. and non-U.S. markets, indicating that the stock selection and portfolio construction methodologies produce superior returns that are not due to chance. The MQ variable, when compared to the average of most models shown in Table 24.3, has a Data Mining Corrections coefficient of 0.47 and is highly statistically significant, having a F-value of 1.872. Thus, one could expect 47% of the excess returns of the MQ model relative to the average return to be continued.

24.4 Overview and Estimation of the 130/30 Strategy

Most investment managers only purchase stocks and seek to outperform the market. These are referred to as long-only managers. Managers who borrow stocks from a broker to sell immediately and repurchase later are called short sellers. Short sellers want the stock price to fall. A relatively recently developed strategy uses 130% of account equity to buy stocks and 30% of account equity to short sell in order to finance the additional stock. Such a portfolio has a net exposure of 100% and is known as a 130/30 portfolio. A 130/30 product not only simultaneously secures both long and short exposures, it also eliminates the long-only constraint required in more traditional investing by allowing short selling.[6] An investment manager decided to secure model exposures that are both long and short. After conducting thorough testing, the quantitative manager made the following observations about a leveraged (130/30) growth portfolio:

- As would be expected for a quantitatively oriented investment process, substantial benefit is missed due to the long-only constraint.
- This portfolio offering used the same investment process and had the same general risk exposures as the manager's other investment management portfolios, that is, exposure to momentum, growth, and selection.
- The portfolio provides both a consistent and long-term risk/return profile that is more attractive than the long-only strategy.

[6] One who buys a stock, or is long a stock, expects the stock's price to rise. Conversely, one who short sells a stock wants that stock's price to fall. For example, a manager short sells a stock by borrowing that stock from a broker, selling it at say, $100, then replaces the broker's stock by repurchasing it later at a lower price, say $75. As a short seller, that manager earns the difference of $25 ($100-$75). However, the manager loses money if the shorted stock rises instead of falls in value.

S. Park and S. Bucknor of CRA RogersCasey attribute the growth of long-short strategies – specifically the 120/20 wave of products (similar to the 130/30 strategy but with a different ratio) – to the ability of quantitative firms and traditional asset managers alike to forecast expected returns for most, if not all, securities in their investable universes.[7]

For various reasons, 130/30 strategies have been gaining traction among investors.[8] A manager typically launches a 130/30 product when the strategy is expected to produce a higher return-to-risk ratio, known as the Sharpe Ratio, than the long-only product, and offers a reasonable tracking error, defined as the standard deviation of the portfolio's return from the benchmark index's return.

This is how the strategy typically works: A 130/30 portfolio invests 130% of the account equity in long stock positions and 30% in short stock positions. The 30% leveraged long positions are financed from the proceeds of the short sales. For example, a $100 million account would hold $130 million of long stocks. The $30 million leveraged portion would be financed through $30 million of short sale proceeds. The net exposure of the portfolio is 100%, or $100 million. As such, a 130/30 portfolio is expected to generate returns of the same order of magnitude as the underlying universe benchmark. Due to this expected return structure and because the net exposure remains at 100%, a 130/30 portfolio is viewed as a substitute for a long-only portfolio allocation. According to Park and Bucknor's survey results, a 130/30-style

[7] 2006, pp. 8–11. Park and Bucknor note that the ability to short securities may improve the overall portfolio return by exploiting unique alpha, or excess return, signals; via a pairs trade; or by reducing industry exposure to the alpha signal. A unique alpha short signal can be the identification of a stock that the manager believes will fall relative to the market. In a pairs trade, the manager identifies two securities, buying one while selling the other, believing that the first stock's price will rise but that the second stock's price will fall, perhaps reverting to a historic ratio. However, an investor need not expect two stocks to move in different directions for a trade; perhaps, the manager is interested in gaining exposure to one company, but not to its industry. Park and Bucknor use the example of the automobile industry, where an investor favors a particular firm, such as Chrysler, although the industry's fundamental data have been weak. The concept of short selling is not new. In 1934, Graham and Dodd wrote in their now-classic *Security Analysis* text that short selling was practiced in the early twentieth century to hedge investments. They also thought that new stock legislation of the 1930s might prevent short selling (p. 284). Despite this legislation, however, short selling has continued to be permitted. Park and Bucknor concur with managers that the 120/20 portfolios are an extension of long-only strategies because these portfolios are measured against the same benchmark.

[8] Philip Middleton says that as much as $50 billion are invested in 130/30 strategies, "130/30 Portfolios – Here Comes the Wave," Merrill Lynch, Global Asset Managers, Industry Overview, March 6, 2007. Alford (2006) attributes particular interest in 130/30 strategies to the following: (1) the desire to hold constant the information ratio, or the relative value of annualized expected (excess) residual return to residual risk (the higher targeted levels of return to risk require more shorting); (2) when equity risk is low, as currently is the case, greater leverage is needed to overcome the loss due to the no-shorting constraint; and (3) managers who specialize in researching overvalued stocks benefit more from shorting than managers who only research stocks to purchase. Smaller managers derive a larger benefit from shorting stocks than managers who concentrate on larger stocks.

long-short portfolio is positioned, by plan sponsors, in the normal investment categories rather than in the alternative investment, or absolute return bucket.[9]

Within broad parameters, a manager with forecasting skill should be able to use a 130/30 format to increase expected returns compared to a long-only offering. When one quantitative manager tested a simulated 130/30 portfolio against its non-U.S. developed growth (long-only) portfolio, the simulated portfolio had a higher Sharpe Ratio than the long-only portfolio.[10] Thus, the return-to-risk ratio was higher for the leveraged portfolio than for the long-only portfolio. A 130/30 portfolio is characterized by higher tracking errors and total risk, as measured by the portfolio's standard deviation, but the investor is compensated for bearing these risks in the higher Sharpe Ratio.

For legal and technical reasons, 130/30 accounts are often held in onshore or offshore limited partnerships or funds. There are audit, legal, and service fees associated with these formats. Even if the 130/30 accounts are held on a separate account basis, there are fees associated with establishing and maintaining prime brokerage accounts (as traditional custodian banks generally cannot hold short positions). Finally, management fees are often higher for 130/30 accounts than for long-only accounts. These higher fees compensate the manager for managing more assets, both long and short, as well as for the generally greater degree of complexity of such portfolios. A prudent investor should understand and accept the higher risks and costs before undertaking a 130/30 investment.

24.5 Experiences with a 130/30 Model

In January 2007, a quantitative manager began managing a 130/30 portfolio in the non-U.S. developed growth space, having limited the universe to the developed

[9] A 130/30 portfolio is an active strategy that can be used by a skilled manager to increase expected portfolio return. For example, suppose that a particular equity universe benchmark has an expected return of 10% and that a manager with skill in that universe has been able to generate excess returns of 5% on average each year, then, with a normal long-only allocation, that manager would, therefore, reasonably be expected to return 15%. Naturally, the actual realized return would depend on the benchmark return in the specific year as well as on the actual excess return generated. Now, consider the expected results using the 130/30 format. Since the net equity exposure for such a portfolio remains at 100%, the benchmark component of the expected return continues to be 10%. However, since the manager is expected to generate excess return of 5% on an unleveraged long portfolio, it would be reasonable to assume excess return of 6.5% on a portfolio with 1.3:1 long-side leverage. If the manager has equal skill on the short side, additional excess return of 1.5% could be expected on the 30% short positions. Accordingly, the total expected return for the 130/30 portfolio would be 18% (the 10% market return plus 5% excess return of stocks plus the leverage factor times the excess returns) compared to 15% using the long-only strategy. Thus, a manager with stock-selection skill could earn higher returns at the same magnitude of market-based or beta risk. Investors should note that market risk is not the sole source of risk; indeed, there are many other sources of risk to consider.

[10] "An Assessment of a McKinley Capital Non-U.S. Developed Growth Portfolio – A Review of the Factors, Returns, and Implementation," August 2006.

markets (EAFE countries plus Canada) because of the potential difficulties and risks involved with shorting emerging market stocks.

The 130/30 portfolio features risk management that uses mathematical techniques to neutralize unintended exposures. By way of comparison, risk management in the long-only portfolio typically relies on more traditional concepts such as sector, country, and industry balance, as well as equal active weighting. However, more sophisticated risk control is required in the 130/30 format due to the need to balance the exposures created by the longs and shorts. The tracking error of the 130/30 portfolio is estimated to be materially higher than the expected tracking error of the long-only portfolio, but consistent with the risk–return tradeoff.

Prior to launching its 130/30 product, the quantitative manager compared the results of its non-U.S. developed growth portfolio with a simulated 130/30 portfolio. The APT Risk Model discussed previously was used to create a market-like 130/30 portfolio using the Blin and Bender risk-control methodology. That is, 20 sources of systematic risk were estimated. [Risk control involves the estimation, using principal components analysis, of at least 20 orthogonal (independent) betas. The APT methodology serves the same purpose as balancing sector, industry, country, size and position, but provided greater control of systematic influences.] Furthermore, the portfolio simulations were built using lambda, a measure of the risk–return tradeoff, of 200, a value consistent with maximizing the Sharpe Ratio. The estimated trading errors exceed bounds of a reasonable, long-run portfolio 6–8% tracking errors. Most models produced excess returns. The results reported in Table 24.6 are consistent with the MQ analysis published in a November 2006 white paper. Generally, those research efforts suggest that relaxing the long-only constraint for a quantitatively oriented manager, even by a marginal degree, leads to a better transfer coefficient, which measures how well an actual portfolio captures a model signal. This, in turn, leads to a better risk/return tradeoff than would otherwise be generated by a long-only portfolio.

The MQ process dominated the EAFE plus Canada universe of models in the 130/30 portfolio construction process. This is a strategy that involves the construction of a long portfolio tilted to securities with positive earnings revisions and positive risk-adjusted relative returns while simultaneously building a short portfolio with the opposite characteristics. This concept allowed the investment manager to test the entire quantitative process within the confines of a single portfolio output. The Sharpe Ratio increased substantially in the 130/30 strategy relative to the corresponding long-only product.[11] That simulation confirmed the manager's risk-and-return expectations for the product. The most important result is when the long-only non-U.S. MQ Sharpe ratio of 0.48 is compared with the corresponding 130/30 Sharpe Ratio of 0.96 for a lambda value of 200, the economic and statistical

[11] A calculation of the performance of the 130/30 MQ strategy developed in this section and subsequent measurement of its performance during the January 2007–May 2008 period shows that the 130/30 portfolio (gross return) outperformed its benchmark, EAFE Growth, 15.08% versus 10.24%, had a higher Sharpe Ratio than its benchmark, 0.63 versus 0.49, and produced an information ratio of 0.64. Many quantitative processes had severe performance issues in this period, notably August 2007 [Khandani and Lo (2007)].

Table 24.6 Data mining corrections models (EAFE + Canada 130/30 strategy, liquidity-screened universe 10/1995–4/2008)

Variable	GM	STD	ShR	TError	IR
MQ	23.22	19.04	1.02	11.35	1.34
CP	16.38	17.37	0.72	8.85	0.62
FEP1	15.74	17.43	0.68	10.52	0.50
FEP2	15.74	17.43	0.68	10.52	0.50
SP	14.78	21.26	0.52	12.07	0.45
PM	13.65	19.03	0.51	11.09	0.51
EP	13.12	16.36	0.57	8.98	0.32
RV2	10.42	16.50	0.40	6.22	0.06
BP	10.23	21.99	0.29	12.61	−0.06
RV1	9.14	16.16	0.33	5.97	−0.05
Benchmark	8.12	14.26	0.30		
FGRI1	7.64	17.07	0.22	8.60	−0.24
FGRI2	5.87	18.56	0.11	8.84	−0.39

GM = geometric mean, *STD* = annualized standard deviation, *ShR* = Sharpe ratio, *TError* = tracking error, IR = information ratio, *MQ* = McKinley quant score, *PM* = price momentum, *BP* = book value/price, CP = cash flow/price, *EP* = earnings/price, *SP* = sales/price, *FEP1* = forecasted FY1 EPS/price, *FEP2* = forecasted FY2 EPS/price, *RV1* = revisions to FY1 EPS, *RV2* = revisions to FY2 EPS, *CTEF* = average (revisions, forecasted EPS yields, breadth), *LTG* = projected long-term EPS forecast, *FGRI1* = FY1 EPS forecast/last year's reported EPS, *FGRI2* = FY2 EPS forecast/last year's reported EPS

dominance of the 130/30 strategy versus the long-only strategy is obvious. The real-time portfolio performance bears out this result for the period through may 2008.

The MQ variable passes the Data Mining Corrections test criteria for both U.S. and non-U.S. markets, indicating that the stock selection and portfolio construction methodologies produce superior returns that are not due to chance. The MQ variable, when compared to the average of all models shown in Table 24.6, has a Data Mining Corrections coefficient of 0.624 and is highly statistically significant, having an F-value of 2.663. Thus, one could expect 62% of the excess returns of the MQ model relative to the average return to be continued. The application of the 130/30 methodology allows the manager to have greater confidence in the continued outperformance than the long-only manager. The application of the Markowitz-Xu Data Mining Corrections test is most appropriate for portfolio construction, management, and evaluation.

Where should an investment manager operate along the efficient frontier? The geometric mean and Sharpe Ratio of the MQ strategy is maximized with a lambda value of 200. The information coefficient is maximized at a lambda of 100 (see Table 24.7). Thus, an investment manager could operate at a lambda value of 100 or 200.

There is an enhancement of approximately 200 basis points in returns in the 130/30 portfolios created with the tracking-error-at-risk methodology compared with the mean-variance approach. This difference was not present in the long-only portfolio construction process.

Table 24.7 MQ 130/30 trade-off analysis (EAFE + Canada strategy 10/1995–4/2008)

Variable	Lambda	GM	STD	ShR	TError	IR
MQ	1000	21.2	21.2	0.82	14.5	0.92
MQ	500	21.4	20.9	0.84	13.5	0.99
MQ	200	22.5	19.5	0.96	11.5	1.27
MQ	100	21.4	18.1	0.96	10.0	1.32
MQ	75	20.1	17.6	0.92	9.3	1.27
MQ	50	18.0	17.5	0.81	8.4	1.13
Benchmark		8.1	14.3	0.30		

GM = geometric mean, *STD* = annualized standard deviation, *ShR* = Sharpe ratio, *TError* = tracking error, IR = information ratio, *MQ* = McKinley quant score

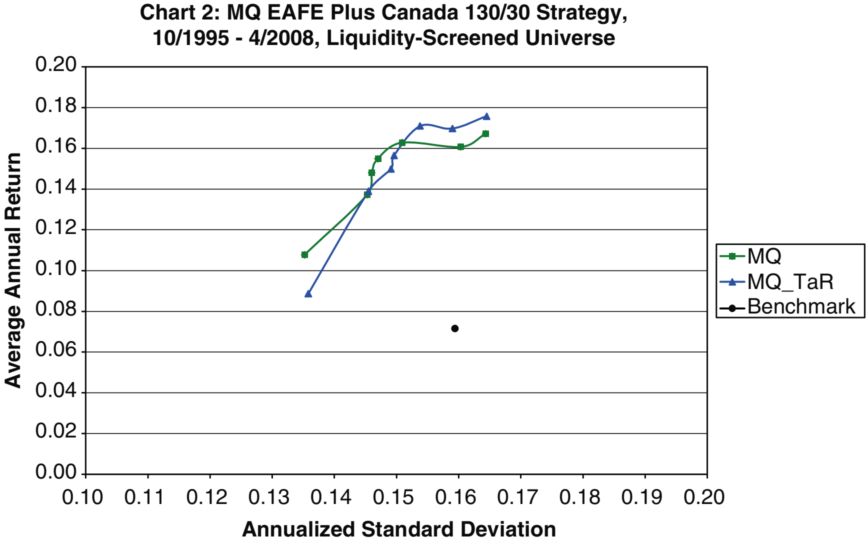


Chart 24.2 MQ EAFE plus Canada 130/30 strategy 10/1995–4/2008, liquidity-screened Universe

Chart 24.2 indicates that the MQ 130/30 strategy can offer a superior investment opportunity set relative to the long-only investment strategy.

24.6 Summary and Conclusion

Quantitative modeling can be effectively used to enhance client wealth. Multi-factor risk models, and APT models in particular, can be used to aid in stock selection and aid the portfolio manager in portfolio construction. An appropriate risk model allows the stock-selection variables with statistically significant ICs to be employed with portfolio construction techniques to produce higher geometric means and Sharpe Ratios than the respective universe benchmarks. The use of ICs may not produce results similar to the asset selection because the market effect and extra-market covariances must be removed to properly estimate the contribution of a variable to the creation of long-only and 130/30 efficient portfolios. The APT framework is useful in portfolio construction of equal active weighting and tracking-error-at-risk portfolios. The use of leverage in a 130/30 context can aid the investment manager in dominating a long-only portfolio manager. Financial variables have been tested and model portfolios have been created that produced returns that pass the Data Mining Corrections tests, and have, initially, produced meaningful excess returns to investors. Moreover, the application of the Markowitz-Xu Data Mining Corrections test to the 130/30 returns produced a higher beta coefficient and level of statistical significance than the long-only results, indicating that the excess returns are not solely due to chance.

Disclaimer

The information contained herein is believed to be acquired from reliable sources but accuracy cannot be guaranteed. This presentation is for informational purposes only; was prepared for academics and financially sophisticated audiences; and is not intended as representing specific financial services or recommendations for specific investment purposes. Furthermore, neither Mr. Guerard nor McKinley Capital Management, Inc., ("McKinley Capital") makes any representations or warranties as to the appropriateness or merit of this analysis for individual use. Investors should seek individualized professional financial advice before investing. Comments and projections are based on information available at the time of writing, and are subject to change without notice.

Past returns are not indicative of future results. All returns are gross of all fees, commissions and investment related costs. Therefore, returns shown herein must be reduced to reflect such charges. International investing involves special risks, including greater economic, political and currency fluctuation risks which may be more pronounced in emerging markets. Shorting stocks also carry additional risks and are more costly to maintain than long positions. Furthermore, standard U.S. GAAP accounting standards may differ from other country accounting principles and thus returns may be higher or lower than anticipated.

Individual security names used herein are for illustrative purposes only, and should not be construed as current or past holdings or recommendation. References to specific indices are for comparison purposes only and it should not be assumed that the authors of this document are recommending or supporting one index over any others. The authors and/or their employer may use and/or rely on specific index names, other financial data and certain analysis without the infringement of copyright materials. However, recipients of this information may not assume that those same rights are transferable. Therefore, redistribution in whole or in part of this presentation is not permitted without McKinley Capital's written authorization. Please refer to each index's web site for complete information about that portfolio. As a standard, individuals cannot directly invest in an index. McKinley Capital has not presented any firm specific performance or material in the document.

To receive the firm's Form ADV II, or a description of all McKinley Capital GIPS® verified composites, please contact us at (907) 563–4488 or visit our website, www.mckinleycapital.com. A bibliography is also provided for easy reference to additional and other quoted materials.

References

Alford, A. 2006. "Demystifying the Newest Equity Long-Short Strategies: Making the Unconventional Conventional." Goldman Sachs Asset Management, October.

Arnott, R. 1985. "The Use and Misuse of Consensus Earnings." *Journal of Portfolio Management*, 18–27.

Basu, S. 1977. "Investment Performance of Common Stocks in Relations to their Price Earnings Ratios: A Test of Market Efficiency." *Journal of Finance* 32, 663–682.

Basu, S. 1983. "The Relationship between Earnings Yield, Market Value and Return for NYSE Common Stocks: Further Evidence." *Journal of Financial Economics* 12, 129–156.

Bernhard, A. 1959. *The Evaluation of Common Stocks*. New York: Simon and Schuster.

Black, F., M.C. Jensen, and M. Scholes. 1972. "The Capital Asset Pricing Model: Some Empirical Tests." In *Studies in the Theory of Capital Markets*, edited by M. Jensen. New York: Praeger.

Blin, J.-M., S. Bender, and J.B. Guerard. 1995. "Why Your Next Japanese Portfolio should be Market Neutral." *Journal of Investing* 4, 33–40.

Blin, J.-M., S. Bender, and J.B. Guerard Jr. 1997. "Earnings Forecasts, Revisions and Momentum in the Estimation of Efficient Market-Neutral Japanese and U.S. Portfolios." In *Research in Finance* 15, edited by A. Chen, Greenwich, CT-JAI Press, Inc.

Bloch, M., J.B. Guerard Jr., H.M. Markowitz, P. Todd, and G. L. Xu. 1993. "A Comparison of Some Aspects of the U.S. and Japanese Equity Markets." *Japan & the World Economy* 5, 3–26.

Brown, S.J. 2008. "Elusive Return Predictability: Discussion," *International Journal of Forecasting* 24, 19–21.

Callen Associates, "What the Heck is 130/30?" Working Paper.

Chan, L.K.C., Y. Hamao, and J. Lakonishok. 1991. "Fundamentals and Stock Returns in Japan." *Journal of Finance* 46, 1739–1764.

Chan, L.K.C., N. Jegadeesh, and J. Lakonishok. 1996. "Momentum Strategies." *Journal of Finance* 51, 1681–1713.

Chan, L.K.C., N. Jegadeesh, and J. Lakonishok. 1999. "The Profitability of Momentum Strategies." *Financial Analysts Journal*, 99, 80–90.

Chen, N.F., R. Roll, and S. Ross. 1986. "Economic Forces and the Stock Market." *Journal of Business* 59, 383–403.

Clarke, R.G., H. de Silva, and S. Thorley, 2002. "Portfolio Constraints and the Fundamental Law of Active Management." *Financial Analyst Journal*, 58, 48–66.

Clarke, R.G., H. de Silva, and S. Sapra, 2004. "Toward More Information-Efficient Portfolios." *The Journal of Portfolio Management*, Fall, 54–63.

Cohen, K.J. and J.A. Pogue. 1967. "An Empirical Evaluation of Alternative Portfolio-Selection Models." *Journal of Business* 40, 166–193.

Conner, G. and R.A. Korajczyk. 1988. "Risk and Return in an Equilibrium APT: Application of a New Test Methodology." *Journal of Financial Economics* 21, 255–289.

Conner, G. and R.A. Korajczyk. 1993. "A Test for the Number of Factors in an Approximate Factor Model." *Journal of Finance* 48, 1263–1291.

Conner, G. and R.A. Korajczyk. 1995. “The Arbitrary Theory and Multifactor Models of Asset Returns.” In *Handbooks in Operations Research and Management Science: Finance* 9, edited by R.A. Jarrow, V. Malsimovic, and W.T. Ziemba, Amsterdam: Elsevier, pp. 87–144.
Dhrymes, P.J., I. Friend, and N.B. Gültekin. 1984. “A Critical Reexamination of the Empirical Evidence on the Arbitrage Pricing Theory.” *Journal of Finance* 39, 323–346.
Dhrymes, P.J., I. Friend, N.B. Gültekin, and M.N. Gültekin. 1985. “An Empirical Examination of the Implications of Arbitrage Pricing Theory.” *Journal of Banking and Finance* 9, 73–99.
Dhrymes, P.J., I. Friend, M.N. Gültekin, and N.B. Gültekin. 1985. “New Tests of the APT and Their Implications.” *Journal of Finance* 40, 659–674.
Dimson, E. 1988. *Stock Market Anomalies*. Cambridge: Cambridge University Press.
Dremen, D. 1979. *Contrarian Investment Strategy*. New York: Random House.
Dremen, D. 1998. *Contrarian Investment Strategies: The Next Generation*. New York: Simon and Schuster.
Elton, E.J. and M.J. Gruber. 1970. “Homogeneous Groups and the Testing of Economic Hypothesis.” *Journal of Financial and Quantitative Analysis* 5, 581–602.
Elton, E.J. and M.J. Gruber. 1973. “Estimating the Dependence Structure of Share Prices – Implications for Portfolio Selection.” *Journal of Finance* 28, 1203–1232.
Elton, E.J., M.J. Gruber, and M.W. Padberg. 1979. “Simple Criteria for Optimal Portfolio Selection: The Multi-Index Case.” In *Portfolio Theory, 25 Years After: Essays in Honor of Harry Markowitz,* edited by E.J. Elton and M.J. Gruber. Amsterdam: North-Holland.
Elton, E.J., M.J. Gruber, and M.N. Gültekin. 1981. “Expectations and Share Prices.” *Management Science* 27, 975–987.
Elton, E.J., M.J. Gruber, and C. Blake. 1995. “Fundamental Economic Variables, Expected Returns, and Bond Fund Performance.” *Journal of Finance* 50, 1229–1256.
Elton, E.J., M.J. Gruber, S.J. Brown, and W.N. Goetzman. 2007. *Modern Portfolio Theory and Investment Analysis, seventh edition*. New York: Wiley.
Fama, E.F. and J.D. MacBeth. 1973. “Risk, Return, and Equilibrium: Empirical Tests.” *Journal of Political Economy* 81, 607–636.
Fama, E.F. 1976. *Foundations of Finance*. New York: Basic.
Fama, E.F. and K.R. French. 1992 “Cross-Sectional Variation in Expected Stock Returns.” *Journal of Finance* 47, 427–465.
Fama, E.F., K.R. French. 1995. “Size and the Book-to-Market Factors in Earnings and Returns. *Journal of Finance* 50, 131–155.
Farrell, J.L. Jr. 1974. “Analyzing Covariance of Returns to Determine Homogeneous Stock Groupings.” *Journal of Business* 47, 186–207.
Farrell, J.L. Jr. 1997. *Portfolio Management: Theory and Applications*. New York McGraw-Hill/Irwin.
Ferson, W.E. and C.R. Harvey. 1991a.“The Variation of Economic Risk Premiums.” *Journal of Political Economy* 99, 385–415.
Ferson, W.E. and C.R. Harvey. 1991b. “Sources of Predictability in Portfolio Returns.” *Financial Analysts Journal* 47, 49–56.
Ferson, W.E. and C.R. Harvey. 1995. “Explaining the Predictability of Asset Returns.” In *Research in Finance* 11, edited by A. Chen, Greenwich, CT-JAI Press, Inc.
Ferson, W.E. 1995. “Theory and Empirical Testing of Asset Pricing Models.” In Chapter 5, *Finance, Handbooks in Operations Research and Management Science*, 9, edited by R. Jarrow, V. Maksimovic, and W.T. Ziemba. Amsterdam: North Holland.
Ferson, W.E. and C.R. Harvey. 1998. “Fundamental Determinants of National Equity Market Returns: A Perspective on Conditional Asset Pricing.” *Journal of Banking & Finance* 21, 1625–1665.
Fogler, R., K. John, and J. Tipton. 1981. “Three Factors, Interest Rate Differentials, and Stock Groups.” *Journal of Finance* 36, 323–335.
Graham, B., and D. Dodd. 1934. *Security Analysis: Principles and Technique*. New York: McGraw-Hill.
Grinhold, R. and R. Kahn. 1999. *Active Portfolio Management* New York: McGraw-Hill/Irwin.

Guerard, J.B. Jr., M. Gültekin, and B.K. Stone. 1997. "The Role of Fundamental Data and Analysts' Earnings Breadth, Forecasts, and Revisions in the Creation of Efficient Portfolios." In *Research in Finance* 15, edited by A. Chen, Greenwich, CT-JAI Press, Inc.

Guerard J.B. Jr. and A. Mark. 2003. "The Optimization of Efficient Portfolios: The Case for a Quadratic R&D Term." In *Research in Finance* 20, Greenwich, CT-JAI Press, Inc. 213–247.

Guerard, J.B. Jr., M. Takano, and Y. Yamane. 1993. "The Development of Efficient Portfolios in Japan with Particular Emphasis on Sales and Earnings Forecasting." *Annals of Operations Research* 45, 91–108.

Haugen, R.A. and N. Baker. 1996. "Communality in the Determinants of Expected Results." *Journal of Financial Economics* 41, 401–440.

Haugen, R.A. 1999. *The Inefficient Stock Market*. Upper Saddle River, NJ: Prentice Hall.

Haugen R.A. 2001. *Modern Investment Theory*, 5th edition. Upper Saddle River, NJ: Prentice Hall.

Jacobs, B.I. and K.N. Levy, 2006. "Enhanced Active Equity Strategies." *Journal of Portfolio Management* Spring, 45–55.

Jacobs, B.I., K.N. Levy, and H.M. Markowitz, 2006. "Trimability and Fast Optimization of Long-Short Portfolios." *Financial Analysts Journal* 62, 36–46.

Jacobs, B.I., K.N. Levy, and H.M. Markowitz, 2006. "Portfolio Optimization with Factors, Scenarios, and Realistic Short Positions", *Operations Research* 53 (2005), 586–599.

Jegadeesh, N. and S. Titman. 1993. "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency." *Journal of Finance* 48, 65–91.

Khandani, A. and A. Lo. 2007. "What Happened to the Quants in August 2007?" MIT Working Paper. http://www.ssrn.com.

Khuri, A.I. and J.A. Cornell. 1996. *Response Surfaces: Designs and Analysis*, New York: Marcel Dekker.

King, B.F. 1966. "Market and Industry Factors in Stock Price Behavior." *Journal of Business* 39, 139–191.

Lakonishok, J., A. Shleifer, and R.W. Vishny. 1994. "Contrarian Investment, Extrapolation, and Risk." *Journal of Finance* 49, 1541–1578.

Latane, H.A., D.L. Tuttle, and C.P. Jones. 1975. *Security Analysis and Portfolio Management*, 2nd Edition. New York: Roland.

Lintner, J. 1965. "The Valuation of Risk Assets on the Selection of Risky Investments in Stock Portfolios and Capital Investments." *The Review of Economics and Statistics*, 13–37.

Markowitz, H.M. 1952. "Portfolio Selection." *Journal of Finance* 7, 77–91.

Markowitz, H.M. 1959. *Portfolio Selection: Efficient Diversification of Investment*. Cowles Foundation Monograph No. 16. New York: Wiley.

Markowitz, H.M. 1976. "Investment in the Long Run: New Evidence for an Old Rule." *Journal of Finance* 31, 1273–1286.

Markowitz, H.M. 1984. "The Two-Beta Trap." *Journal of Portfolio Management* 11, 12–19.

Markowitz, H.M. 1987. *Mean-Variance Analysis in Portfolio Choice and Capital Markets*. Oxford: Basil Blackwell.

Markowitz, H.M. and G. Xu. 1994. "Data Mining Corrections." *Journal of Portfolio Management* 21, 60–69.

Markowitz, H.M. 2000. *Mean-Variance Analysis in Portfolio Choice and Capital Markets*. New Hope, PA: Frank J. Fabozzi.

McKinley Capital Management, 2006. "An Assessment of a McKinley Capital Non-U.S. Developed Growth Portfolio – A Review of the Factors, Returns, and Implementation," August.

McKinley Capital Management, 2006. "An Assessment of the McKinley Capital Non-U.S. Developed Growth and Non-U.S. Growth Portfolios," November.

Middleton, P. 2007. "130/30 Portfolios – Here Comes the Wave," Merrill Lynch, Global Asset Managers, Industry Overview, March 6.

Mossin, J. 1966. "Equilibrium in a Capital Asset Market." *Econometrica* 34, 768–783.

Mossin, J. 1973. *Theory of Financial Markets*. Englewood Cliffs, NJ: Prentice-Hall.

Park, S. and S. Bucknor, 2006. "120/20 Portfolios: Long-Only Managers' Way of Reaching for Higher Alpha," CRA RogersCasey, *U.S. Equity Research*, November.

Ramnath, S., S. Rock, and P. Shane. 2008. "The Financial Forecasting Literature: A Taxonomy with Suggestions for Further Research." *International Journal of Forecasting* 24, 34–75.

Reinganum, M. 1981. "The Arbitrage Pricing Theory: Some Empirical Tests." *Journal of Finance* 36, 313–321.

Roll, R. and S.A. Ross. 1984. "The Arbitrage Pricing Theory Approach to Strategic Portfolio Planning." *Financial Analysts Journal* 40, 14–26.

Rosenberg, B. 1974. "Extra-Market Components of Covariance in Security Returns." *Journal of Financial and Quantitative Analysis* 9, 263–274.

Rosenberg, B., M. Hoaglet, V. Marathe, and W. McKibben. 1975. "Components of Covariance in Security Returns." Working paper No. 13, Research Program in Finance. Institute of Business and Economic Research. University of California. Berkeley.

Rosenberg, B. 1976. "Security Appraisal and Unsystematic Risk in Institutional Investment." Proceedings of the Seminar for Research in Security Prices. University of Chicago, Chicago.

Rosenberg, B. and A. Rudd, 1977. "The Yield/Beta/Residual Risk Tradeoff." Working paper No. 66. Research Program in Finance. Institute of Business and Economic Research. University of California, Berkeley.

Rosenberg, B. and V. Marathe, 1979. "Tests of Capital Asset Pricing Hypotheses." In *Research in Finance* 1, 1979, edited by H. Levy, Greenwich, CT-JAI Press, Inc.

Ross, S.A. 1976. "The Arbitrage Theory of Capital Asset Pricing." *Journal of Economic Theory* 13, 341–360.

Ross, S.A. and R. Roll. 1980. "An Empirical Investigation of the Arbitrage Pricing Theory." *Journal of Finance* 35, 1071–1103.

Rudd, A. and B. Rosenberg. 1979. "Realistic Portfolio Optimization." In *Portfolio Theory, 25 Years After*, edited by E. Elton and M. Gruber. Amsterdam: North-Holland.

Rudd, A. and B. Rosenberg. 1980. "The 'Market Model' in Investment Management." *Journal of Finance* 35, 597–607.

Sharpe, W.F. 1963. "A Simplified Model for Portfolio Analysis." *Management Science* 9.

Sharpe, W.F. 1964. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *Journal of Finance* 19, 425–442.

Sharpe, W.F. 1971. "A Linear Programming Approximation for the General Portfolio Analysis Problem." *Journal of Financial and Quantitative Analysis* 6, 1263–1275.

Stone, B.K. 1970. *Risk, Return, and Equilibrium: A General Single-Period Theory of Asset Selection and Capital Market Equilibrium*. Cambridge, MA: MIT.

Stone, B.K. 1973. "A Linear Programming Formulation of the General Portfolio Selection Problem." *Journal of Financial and Quantitative Analysis* 8, 621–636.

Stone, B.K. 1974. "Systematic Interest-Rate Risk in a Two-Index Model of Returns." *Journal of Financial and Quantitative Analysis* 9, 709–721.

Stone, B.K., J.B. Guerard Jr., M. Gültekin, and G. Adams. 2002. "Socially Responsible Investment Screening." Working Paper. Marriott School of Management. Brigham Young University, Provo, UT.

Von Hohenbalken, B. 1975. "A Finite Algorithm to Maximize Certain Pseudoconcave Functions on Polytopes." *Mathematical Programming* 8, 189–206.

Ziemba, W.T. 1992. *Invest Japan*. Chicago: Probus.

Chapter 25
Distortion Risk Measures in Portfolio Optimization

Ekaterina N. Sereda, Efim M. Bronshtein, Svetozar T. Rachev, Frank J. Fabozzi, Wei Sun, and Stoyan V. Stoyanov

25.1 Introduction

The selection of the appropriate portfolio risk measures continues to be a topic of heated discussion and intensive investigations in investment management, as all the proposed risk measures have drawbacks and limited applications. The major focus of researchers[1] has been on the "right" or "ideal" risk measure to be applied in portfolio selection. The principal complexity, however, is that the concept of risk is highly subjective, because every market player has its own perception of risk. Consequently, Balzer (2001) concludes, there is "no single universally acceptable risk measure." He suggests the following features that an investment risk measure should satisfy: relativity of risk, multidimensionality of risk, asymmetry of risk, and nonlinearity.

Rachev et al. (2008) summarize the desirable properties of an "ideal" risk measure, capturing fully the preferences of investors. These properties relate to

E.N. Sereda
HECTOR School of Engineering and Management, University of Karlsruhe, Karlsruhe, Germany

E.M. Bronshtein
Ufa State Aviation Technical University, Ufa, Russian Federation

S.T. Rachev (✉)
University of Karlsruhe, KIT, University of California at Santa Barbara, and FinAnalytica Inc.
e-mail: rachev@kit.edu

S.V. Stoyanov
FinAnalytica, Inc., Seattle, WA, USA

F.J. Fabozzi
Yale School of Management, New Haven, CT, USA

W. Sun
School of Economics and Business Engineering, University of Karlsruhe, Karlsruhe, Germany

[1] See, for example, Goovaerts et al. (1984), Artzner et al. (1999), Kaas et al. (2001), Goovaerts et al. (2003), Zhang and Rachev (2006), Denuit et al. (2006), and Rachev et al. (2008).

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_25,

© Springer Science+Business Media, LLC 2010

investment diversification, computational complexity, multiparameter dependence, asymmetry, nonlinearity, and incompleteness. However, every risk measure proposed in the literature possesses only some of these properties. Consequently, proposed risk measures are insufficient and, based on this, Rachev et al. (2008, p. 4) conclude that an ideal measure does not exist. However, they note that "it is reasonable to search for risk measures which are ideal for the particular problem under investigation."

Historically, the most commonly used risk measure is the standard deviation (variance) of a portfolio's return. In spite of its computational simplicity, variance is not a satisfactory measure due to its symmetry property and inability to consider the risk of low probability events.

A risk measure that has received greater acceptance in practice is value at risk (VaR). Unfortunately, because VaR fails to satisfy the subadditivity property and ignores the potential loss beyond the confidence level, researchers and practitioners[2] have come to realize its limitations, limiting its use for reporting purposes when regulators require it or when a simple to interpret number is required by clients.

A major step in the formulation of a systematic approach toward risk measures was taken by Artzner et al. (1999). They introduced the notion of "coherent" risk measures. It turns out that VaR is not a coherent risk measure. In contrast, a commonly used risk measure in recent years, conditional value at risk (CVaR), developed by Rockafellar et al. (2002), is, in fact, a coherent risk measure. The most general theoretical result about coherent measures is the class of distortion risk measures, introduced by Denneberg (1990) and Wang et al. (1997).

Distortion risk measures were obtained by the simultaneous use of two approaches[3] to define the particular class of risk measures: axiomatic definition and the definition from the economic theory of choice under uncertainty. Due to the second approach, distortion risk measures have their roots in the dual utility theory of Yaari (1987). Using the expected utility's set of axioms with a modified independence axiom, Yaari (1987) developed the distortion utility theory. He has shown that there must exist a "distortion function" such that a prospect is valued at its distorted expectation. Instead of using the tail probabilities in order to quantify risk, the decision maker uses the distorted tail probabilities. For the axiomatic definition, Wang et al. (1997) postulated the axioms to characterize the price of insurance risk. These axioms include the following: law invariance, monotonicity, comonotonic additivity, and continuity. They also proved that risk measures hold such properties if and only if they have the Choquet integral representation with respect to a distorted probability.

Distortion risk measures were originally applied to a wide variety of insurance problems such as the determination of insurance premiums, capital requirement, and capital allocation. Because insurance and investment risks are closely related,

[2] See, for example, Artzner et al. (1999), Szegö (2004), and Zhang and Rachev (2006).

[3] The combination of two approaches – as demonstrated by Föllmer and Schied (2002a), Tsanakas and Desli (2003), and Denuit et al. (2006) – add better perception of the inherent properties of risk measures.

the investment community started to apply distortion risk measures in the context of the asset allocation problem.[4] Wang 2004 has applied the distortion risk measure to price catastrophe bonds and Fabozzi and Tunaru (2008) to price real-estate derivatives.

In the application of portfolio selection, distortion risk measures with the concave distortion function reveal the desired properties, such as law-invariance, subadditivity, and consistency with second-order stochastic dominance. Law-invariance is the prerequisite for the ability to quantify risk of a portfolio from historical data. Subadditivity secures the diversification effect. In general, the motive for constructing a portfolio is to reduce overall investment risk through diversification as set forth by Markowitz (1952), that is, investing in different asset classes and in securities of many issuers. Diversification ensures the avoidance of extreme poor portfolio performance caused by the underperformance of a single security or an industry. The consistency with second-order stochastic dominance provides the link between the construction of risk measures and the decision theory under uncertainty.

In this paper, we propose new distortion risk measures, adding the asymmetric property to the already existing properties of concave distortion risk measures. We do so by extending the Choquet integral construction using quadratic and power distortion functions with different concave parameters in order to better capture the risk perception of investors.

The paper is organized as follows. Section 25.2 starts with the general definition of risk and provides various classifications of risk measures that have appeared in the literature. Sections 25.3 and 25.4 provide a discussion of the distortion risk measures and their properties. In Sect. 25.5, we give examples of distortion functions and show how distortion risk measures are related to VaR and CVaR. We propose the new distortion risk measures with asymmetric property in Sect. 25.6. Section 25.7 summarizes our paper. The appendix contains the properties of risk measures and we make reference to them by property number in the body of the paper.

25.2 Classes of Risk Measures

In general, a risk measure, $\rho: \mathcal{X} \to \mathbb{R}$, is a functional that assigns a numerical value to a random variable representing an uncertain payoff. $\mathcal{X}$ is defined on $L^\infty(\Omega, \mathcal{F}, P)$,[5] the space of all essentially bounded random variables defined on the general probability space $(\Omega, \mathcal{F}, P)$. Not every functional corresponds to the

[4] See, for example, Van der Hoek and Sherris (2001), Gourieroux and Liu (2006), Hamada et al. (2006), and Balbas et al. (2007).

[5] It could be efficient to use unbounded random variables for modeling risks, as far as financial risk has no limits. The implications of risk measure properties can be different on certain finite and on nonatomic probability spaces. See Bäuerle and Müller (2006) and Inoue (2003) for further results on the extension from L^∞ to L^1 probability space.

intuitive notion of risk. One of the main characteristics of such a function is that a more uncertain return should conform to a higher functional value.

Goovaerts et al. (1984) presented the pioneering work of the axiomatic approach to risk measures in actuarial science, where risk measures were analyzed within the framework of premium principles. Artzner et al. (1999) extended the use of this axiomatic approach in the financial literature. The axiomatic definition of how risk is measured includes the setting of the assorted properties (axioms) on a random variable and then the determination of the mathematical functional fitting to the set of axioms.

25.2.1 Pederson and Satchell's Class of Risk Measures

Pederson and Satchell (1998) define risk as a deviation from a location measure. They provided four desirable properties of a "good financial risk measure", such as nonnegativity, positive homogeneity, subadditivity, and translation invariance.[6] Pedersen and Satchell also presented in their work the full characterization of the appropriate risk measures according to their system of axioms.

25.2.2 Coherent Risk Measures

The idea of coherent risk measures was introduced by Artzner et al. (1999). Coherent risk measures are those measures which are translation invariant, monotonous, subadditive, and positively homogeneous.[7] Coherent measures have the following general form:

$$\rho(X) = \sup_{Q \in \mathcal{Q}} E_Q[-X],$$

where $\mathcal{Q}$ is some class of probability measures on Ω.

Four criteria proposed by Artzner et al. (1999) provide rules for selecting and evaluating risk measures. However, one should be aware that not all risk measures satisfying the four proposed axioms are reasonable to use under certain practical situations. Wang (2002) argued that "a risk measure should go beyond coherence" in order to utilize useful information in a large part of a loss distribution. Dhaene et al. (2003), observing "best practice" rules in insurance, concluded that coherent risk measures "lead to problems."

[6] Property 25.8, Property 25.2, Property 25.3.1, and Property 25.6.3, respectively.

[7] Property 25.6.3, Property 25.5, Property 25.3.1, and Property 25.2, respectively.

25.2.3 Convex Risk Measures

Convex risk measures (also called weakly coherent risk measures) were studied by Föllmer and Schied (2002a, b) and Frittelli and Rosazza Gianin (2005). Convex risk measures are a generalization of coherent risk measures obtained by relaxation of the positive homogeneity assumption (Property 25.2) together with the subadditivity condition (Property 25.3.1) and require the weaker property of convexity (Property 25.4). Any convex risk measure takes into account a nonlinear increase of the risk with the size of the position and has the following structure:

$$\rho(X) = \sup_{Q \in \mathcal{Q}} \left(E_Q[-X] - \alpha(Q) \right),$$

where α is a penalty function defined on probability measures on Ω.

Following Frittelli and Rosazza Gianin (2005), a functional $\rho : \mathcal{X} \to \mathbb{R}$ is a convex risk measure if it satisfied convexity (Property 25.4), lower semicontinuity (Property 25.9.5), and normalization ($\rho(0) = 0$) conditions. Bäuerle and Müller (2006) proposed replacing the convexity axiom by the weaker but more intuitive property of consistency with respect to convex order (Property 25.7.6).

25.2.4 Law-Invariant Coherent Risk Measures

Following the notation of Kusouka (2001), law-invariant coherent risk measures have the form:

$$\rho_\alpha(X) \triangleq \frac{1}{\alpha} \int_{1-\alpha}^{1} Z_{-X}(x)\,\mathrm{d}x.$$

where Z: [0,1) → R is non-decreasing and right continuous. This class of risk measures satisfies the lower semicontinuity property (Property 25.9.5) for all $X \in L^\infty$, $0 \le \alpha \le 1$. The class of insurance prices characterized by Wang et al. (1997) is an example of law-invariant coherent risk measures.

25.2.5 Spectral Risk Measures

Spectral measures of risk[8] can be defined by adding two axioms to the set of coherency axioms: law invariance (Property 25.1) and comonotonic additivity

[8] See Kusouka (2001), Acerbi (2002), and Adam et al. (2007).

(Property 25.3.2). Spectral risk measures consist of a weighted average of the quantiles of the returns distribution using a nonincreasing weight function[9] referred to as a spectrum and denoted by ϕ. It is defined as follows:

$$M_{\phi}(X) = -\int_0^1 \phi(x) F_X(x)\,\mathrm{d}x,$$

where ϕ is a nonnegative, nonincreasing, right-continuous integrable function defined on $[0, 1]$ and such that $\int_0^1 \phi(x)\,\mathrm{d}x = 1$. Assumptions made on ϕ determine the coherency of spectral risk measures. If any of these assumptions is relaxed, the measure is no longer coherent. Spectral risk measures possess positive homogeneity (Property 25.2), translation invariance (Property 25.6.3), monotonicity (Property 25.5), subadditivity (Property 25.3.1), law invariance (Property 25.1), comonotonic additivity (Property 25.3.2), consistency with second-order stochastic dominance (SSD) (Property 25.7.4), and expected utility theory.

25.2.6 Deviation Measures

Rockafellar et al. (2002)[10] defined deviation measures as positive, subadditive, positively homogeneous, Gaivoronsky-Pflug (G–P) translation invariant[11] risk measures. Deviation measures are normally used by totally risk-averse investors.

25.2.7 Expectation-Bounded Risk Measures

Rockafellar et al. (2002) proposed expectation-bounded risk measures, imposing the conditions of subadditivity, positive homogeneity, translation invariance, and an additional property of expectation-boundedness.[12] There exists a corresponding one-to-one relationship between deviation measures and expectation-bounded risk measures. One can derive expectation-bounded coherent risk measures if additionally monotonicity (Property 25.5) is satisfied.

[9] ϕ can be observed as a weighted function reflecting an investor's subjective risk aversion.

[10] See also Rockafellar et al. (2003, 2006).

[11] Property 25.8, Property 25.3.1, Property 25.2, and Property 25.6.2, respectively.

[12] Property 25.3.1, Property 25.2, Property 25.6.3, and Property 25.10, respectively.

25.2.8 Reward Measures

De Giorgi (2005) introduced the first axiomatic definition for reward measures and provided their characterization. According to De Giorgi, such measures should satisfy the following conditions: additivity, positive homogeneity, isotonicity with respect to SSD, and risk-free condition.[13]

25.2.9 Parametric Classes of Risk Measures

Stone (1973) defined a general three-parameter class of risk measures, which has the form

$$R[c, k, A] = \left(\int_A^{-\infty} |y - c|^k f(y)\,\mathrm{d}y\right)^{1/k},$$

where $A, c \in \mathbb{R}$, and $k > 0$. Stone's class of risk measures includes several commonly used measures of risk and dispersion, such as the standard deviation, the semistandard deviation, and the mean absolute deviation.

Pederson and Satchell (1998) generalized Stone's class of risk measures and introduced the five-parameter class of risk measures:

$$R[A, c, \alpha, \theta, w(\cdot)] = \left[\int_A^{-\infty} |y - c|^\alpha w\left[F(y)\right] f(y)\mathrm{d}y\right]^\theta$$

for some bounded function $w(\cdot)$, $A, c \in \mathbb{R}, \alpha > 0, \theta > 0$. This class of risk measures also includes the lower partial moments as an extension of the Stone class. Ebert (2005) argues that because of the confusing number of parameters presented by Pedersen and Satchell, "it seems to be impossible to comprehend their meaning and their interaction."

25.2.10 Quantile-Based Risk Measures

Quantile-based risk measures include value at risk, expected shortfall, tail conditional expectation, and worst conditional expectation. We describe each measure below.

Value at risk (VaR) specifies how much one can lose with a given probability (confidence level). Its formal definition is

$$\mathrm{VaR}^\alpha(X) = -x^{(\alpha)} = q_{1-\alpha}(-X).$$

[13] Property 25.3.2, Property 25.2, Property 25.7.4, and Property 25.13.2, respectively.

VaR has the following properties: monotonicity (Property 25.5), positive homogeneity (Property 25.2), translation invariance (Property 25.6.3), law invariance (Property 25.1), and comonotonic additivity (Property 25.3.2). VaR possesses the subadditivity attribute (Property 25.3.1) only for joint-elliptically distributed risks (see Embrechts et al. 2002), but this assumption is rare in practice.

Despite its simplicity and wide applicability, VaR is controversial. A common criticism among academics is that VaR is not subadditive, hence not coherent and that VaR calculations lead to substantial estimation errors (see, e.g., Artzner et al. 1999; Szegö 2004; and Zhang and Rachev 2006). A risk manager should be aware of its limitations and use it properly.

Expected shortfall (ES), also known as tail (or conditional) VaR (see Rockafellar et al. 2002), corresponds to the average of all VaRs above the threshold α:

$$\mathrm{ES}^{\alpha}(X) = \int_{\alpha}^{1} x^{(\alpha)}\,\mathrm{d}x, \quad \alpha \in (0,1).$$

ES was proposed in order to overcome some of the theoretical weaknesses of VaR. ES has the following properties: law invariance (Property 25.1), translation invariance (Property 25.6.3), comonotonic additive (Property 25.3.2), continuity (Property 25.9), monotonicity (Property 25.5), and subadditivity (Property 25.3.1). ES, being coherent,[14] was proposed by Artzner et al. (1999) as a "good" risk measure.

Tail conditional expectation (TCE) was proposed by Artzner et al. (1999) in the following form:

$$\mathrm{TCE}^{\alpha}(X) = -E\{X|X \le x^{(\alpha)}\}.$$

TCE^{α} does not possess the subadditivity property for general distributions; it is coherent only for continuous distributions.

Worst conditional expectation (WCE) is defined as

$$\mathrm{WCE}_{\alpha}(X) = -\inf\{E[X|A]: A \in \mathcal{F}, P(A) > \alpha\}.$$

$\mathrm{WCE}_{\alpha}(X)$ is not law invariant, so it cannot be estimated solely from data. Using such measures can lead to different risk values for two portfolios with identical loss distributions.

Comparing ES, TCE, and WCE, one finds that

$$\mathrm{TCE}^{\alpha}(X) \le \mathrm{WCE}_{\alpha}(X) \le \mathrm{ES}_{\alpha}(X).$$

ES has the maximum value among TCE and WCE when the underlying probability space varies. If the distribution of X is continuous, then

$$\mathrm{TCE}^{\alpha}(X) = \mathrm{WCE}_{\alpha}(X) = \mathrm{ES}_{\alpha}(X).$$

[14] See Acerbi and Tasche (2004).

25.2.11 Drawdown Measures

Drawdown measures are intuitive measures. A psychological issue in handling risk is the tendency of people to compare the current situation with the very best one from the past. Drawdowns measure the difference between two observable quantities – local maximum and local minimum of the portfolio wealth. Cheklov et al. (2003) defined the drawdown function as the difference between the maximum of the total portfolio return up to time t and the portfolio value at t.

Drawdown measures are close to the notion of deviation measure. Examples of drawdown measures include absolute drawdown (AD), maximum drawdown (MDD), average drawdown (AvDD), drawdown at risk (DaR), and conditional drawdown at risk (CDaR). In spite of their computational simplicity, drawdown measures cannot describe the real market situations, and therefore, should be used in combination with other measures.

25.3 Distortion Risk Measures

A *distortion risk measure* can be defined as the distorted expectation of any non-negative loss random variable X. It is accomplished by using a "dual utility" or the distortion function g[15] as follows:

$$\rho_g(X) = \int_0^\infty g(1 - F_X(x))\,\mathrm{d}x = \int_0^1 F_X^{-1}(x)(1-q)\,\mathrm{d}g(q), \tag{25.1}$$

where $g\colon [0,1] \to [0,1]$ is a continuous increasing function with $g(0)=0$ and $g(1) = 1$; $F_X(x)$ denotes the cumulative distribution function of X, while $g(F_X(x))$ is referred to as a distorted distribution function.

For the gain/loss distributions, when the loss random variable can take any real number, $X \in \mathbb{R}$, the distortion risk measure is obtained as follows:

$$\rho_g(X) = \int_0^1 F_X^{-1}(x)dH(x) = -\int_{-\infty}^0 H(F_X(x))\,\mathrm{d}x + \int_0^\infty [1 - H(F_X(x))]\,\mathrm{d}x.$$

where $H(u) = 1 - g(1-u)$. The same holds with the replacement by the survival function $S(x) = 1 - F(x) = P[X > x]$:

$$\rho_g(X) = -\int_{-\infty}^0 [1 - g(S_X(x))]\,\mathrm{d}x + \int_0^\infty g(S_X(x))\,\mathrm{d}x.$$

---

[15] Consider the set function $g\colon \mathcal{F} \to [0,\infty)$, defined on the σ-algebra $\mathcal{F}$, such that $g(\emptyset) = 0$ and $A \subseteq B \Rightarrow g(P[A]) \leq g(P[B])$, for $A, B \in \mathcal{F}$. Such a function g is called a distortion function, and $P[A]$, $P[B]$ are referred to as distorted probabilities.

Van der Hoek and Sherris (2001) developed a more general class of distortion risk measures, depending on the choice of parameters α, g, and h. It has the following form:

$$H_{\alpha,g,h}(X) = \alpha + H_h\left((X-\alpha)^+\right) - H_g\left((\alpha - X)^+\right),$$

where $\alpha^+ = \max[0, \alpha]$. When $\alpha = 0$ and $h(x) = 1 - g(1-x)$, then we again obtain the Choquet integral representation.

25.4 Properties of Distortion Risk Measures

The properties of the distortion risk measures correspond to the following standard results about the Choquet integral (see Denneberg 1994):

1. If $X \geq 0$, then $\rho_g(X) \geq 0$, monotonicity.
2. $\rho_g(\lambda X) = \lambda \rho_g(X)$, for all $\lambda \geq 0$, positive homogeneity.
3. $\rho_g(X+c) = \rho_g(X) + c$, for all $c \in \mathbb{R}$, translation invariance.[16]
4. $\rho_g(-X) = -\rho_{\tilde{g}}(X)$, where $\tilde{g}(x) = 1 - g(1-x)$.[17]

[16] The proof is as follows:

$$\begin{aligned}\rho_g(X+c) &= \int_0^{-\infty} [g(S_{X+c}(x) - 1]\,\mathrm{d}x + \int_0^c g(S_{X+c}(x))\,\mathrm{d}x + \int_c^\infty g(S_{X+c}(x))\,\mathrm{d}x \\ &= \int_{-\infty}^0 [g(S_X(x-c)) - 1]\,\mathrm{d}x + \int_0^c g(S_X(x-c))\,\mathrm{d}x + \int_c^\infty g(S_X(x-c))\,\mathrm{d}x.\end{aligned}$$

By replacing $x = c + u$, we get

$$\begin{aligned}\rho_g(X+c) &= \int_{-\infty}^{-c} [g(S_X(u)) - 1]\,\mathrm{d}u + \int_{-c}^0 g(S_X(u))\,\mathrm{d}u + \int_0^\infty g(S_X(u))\,\mathrm{d}u \\ &= \int_{-\infty}^0 [g(S_X(u)) - 1]\,\mathrm{d}u + \int_0^\infty g(S_X(u))\,\mathrm{d}u + \int_{-c}^0 \mathrm{d}u \\ &= \rho_g(X) + c.\end{aligned}$$

[17] The proof is as follows:

$$\begin{aligned}\rho_g(-X) &= \int_0^{-\infty} [g(S_{-X}(x) - 1]\,\mathrm{d}x + \int_0^\infty g(S_{-X}(x))\,\mathrm{d}x \\ &= \int_{-\infty}^u [g(1 - S_X(-x) + P[X=x]) - 1]\,\mathrm{d}x + \int_{-\infty}^u g(1 - S_X(-x) + P[X=x])\,\mathrm{d}x \\ &= \int_{-\infty}^u (g(1 - S_X(-x)) - 1)\,\mathrm{d}x + \int_{-\infty}^u g(1 - S_X(-x))\,\mathrm{d}x.\end{aligned}$$

Replacing x by $-u$, we get

$$\rho_g(-X) = -\int_{-\infty}^u (g(1 - S_X(u)) - 1)\,\mathrm{d}u - \int_{-\infty}^u g(1 - S_X(u))\,\mathrm{d}u = -\rho_{\tilde{g}}(X).$$

5. If a random variable X_n has a finite number of values (i.e., $X_n \xrightarrow{w} X$) and $\rho_g(X)$ exists, then $\rho_g(X_n) \to \rho_g(X)$. This property implies that it is enough to prove the statement for the discrete random variables, and then carry over the result to the general continuous case.
6. If X and Y are comonotonic risks, taking positive and negative values, then

$$\rho_g(X+Y) = \rho_g(X) + \rho_g(Y)0$$

   In the literature, this property is called *comonotonic additivity.*
7. In the generalized case, distortion risk measures are not additive[18]:

$$\rho_g(X+Y) \neq \rho_g(X) + \rho_g(Y).$$

8. Distortion risk measures are subadditive if and only if the distortion function $g(x)$ is concave.

$$\rho_g(X+Y) \leq \rho_g(X) + \rho_g(Y).$$

   The proof is given in Wirch and Hardy (1999). Hence, concave distortion risk measures are coherent risk measures.
9. For a nondecreasing distortion function g, the associated risk measure ρ_g is consistent with the stochastic dominance of order 1

$$X \leq_1 Y \Rightarrow \rho_g(X) \leq \rho_g(Y).$$

   The proof is given in Hardy and Wirch (2003).
10. For a nondecreasing concave distortion function g, the associated risk measure ρ_g is consistent with the stochastic dominance of order 2 (i.e., SSD)

$$X \leq_2 Y \Rightarrow \rho_g(X) \leq \rho_g(Y).$$

   As a result, every coherent distortion risk measure is consistent with respect to SSD.
11. For a strictly concave distortion function g, the associated risk measure ρ_g is strictly consistent with SSD

$$X <_2 Y \Rightarrow \rho_g(X) < \rho_g(Y).$$

   The proof is given in Hardy and Wirch (2003).

[18] The proof is as follows. Consider the function $g = x^2$; the joint distribution of discrete random variables X and Y is defined as follows: $P(1,1) = P(-1,1) = P(1,-1) = P(-1,-1) = 0.25$. The marginal distributions X and Y have the forms: $P(1) = P(-1) = 0.5$. Direct calculations show that

$$\rho_g(X) = \rho_g(Y) = -0.5; \quad \rho_g(X+Y) = 0.25; \quad 0.25 \neq -1$$

The risks X and Y are independent here.

12. Consistency of distortion risk measures with respect to the higher-order stochastic dominances was analyzed in the financial and actuarial literatures. In particular, Hürlimann (2004) obtained some results about the consistency of distortion risk measures with stochastic dominance of order 3. The necessary precondition to that is the consistency with respect to 3-convex order. The only distortion risk measures which are consistent with 3-convex order are $g(x) = \sqrt{x}$ and $g(x) = x$ under the assumption that the set of possible losses contains all Pareto[19] variables [Theorem 6.3 in Hürlimann (2004)]. Under a much weaker hypothesis of discrete losses, Bellini and Caperdoni (2006) showed that the only coherent distortion risk measure that is consistent with respect to the 3-convex order is the expected value, when $g(x) = x$, leaving the problem open for the case of continuous losses.

25.5 Examples of Distortion Risk Measures

As explained above, the choice of distortion function specifies the distortion risk measures. Thus, finding "good" distorted risk measures boils down to the choice of a "good" distortion function. The properties one might use as a criteria for the choice of a distortion function include *continuity*, *concavity*, and *differentiability*. Many different distortions g have been proposed in the literature. Some well-known ones are presented below. A summary of other proposed distortion functions can be found in Denuit et al. (2005).

- With $g(x) = x$, we have $\rho_g(X) = E[X]$, if the mathematical expectation exists.[20]

[19] With a generic location and scale parameter allowed.

[20] We limit our proof to the interval $[-a, a]$. In this case the mathematical expectation accurately exists.

$$\begin{aligned}
\rho_g(X) &= \int_0^{-a} (S_X(x) - 1)\,\mathrm{d}x + \int_a^0 S_X(x)\,\mathrm{d}x \\
&= -\int_{-a}^0 F_X(x)\,\mathrm{d}x + \int_0^a (1 - F_X(x))\,\mathrm{d}x \\
&= a - \int_{-a}^a F_X(x)\,\mathrm{d}x.
\end{aligned}$$

By integrating by parts, we get

$$\rho_g(X) = a - xF_X(x)|_{-a}^a + \int_{-a}^a x\,\mathrm{d}F_X(x) = E[X],$$

as $F_X(a) = 1$, $F_X(-a) = 0$. In this particular case, the distortion risk measures are additive.

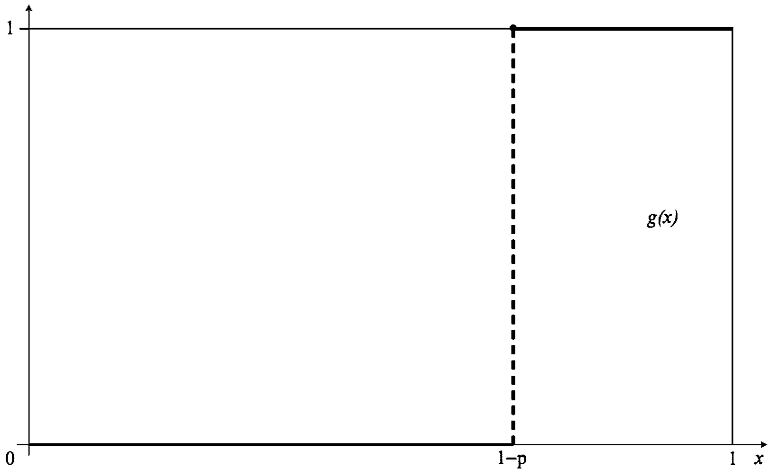


Fig. 25.1 Distortion function of VaR

- VaR corresponds to the distortion:

$$g(x) = \begin{cases} 0, & \text{if } x < 1 - p; \\ 1, & \text{if } x \geq 1 - p. \end{cases}$$

The distortion function is discontinuous in this case due to the jump at $x = 1 - p$ (see Fig. 25.1). This predetermines that VaR is not coherent. As a result, VaR does not represent a "well" behaved distortion function.

- CVaR can be defined as a distortion risk measure based on the distortion function

$$g(x) = \min\left(\frac{x}{1-p}, 1\right), \quad x \in [0, 1].$$

Figure 25.2 presents the given function. It is continuous, implying that CVaR is coherent. But the distortion function of CVaR is not differentiable at $x = 1 - p$. Consequently, it discards potentially valuable information because it maps all percentiles below $(1 - p)$ to a single point "0." By doing so, it fails to take into account the severity of extreme values (Wang 2002).

- In order to overcome these sorts of problems, Wang (2002) considers the following specification of g:

$$g(x) = \Phi\left(\Phi^{-1}(x) - \Phi^{-1}(q)\right),$$

for $p \in [0, 1]$, where $0 < q \leq 0.5$ is some parameter.[21] The distortion function g is indeed nondecreasing, concave, and such that $g(0) = 0$ and $g(1) = 1$.

[21] We need $q \leq 0.5$ in order to get concave distortion risk measures.

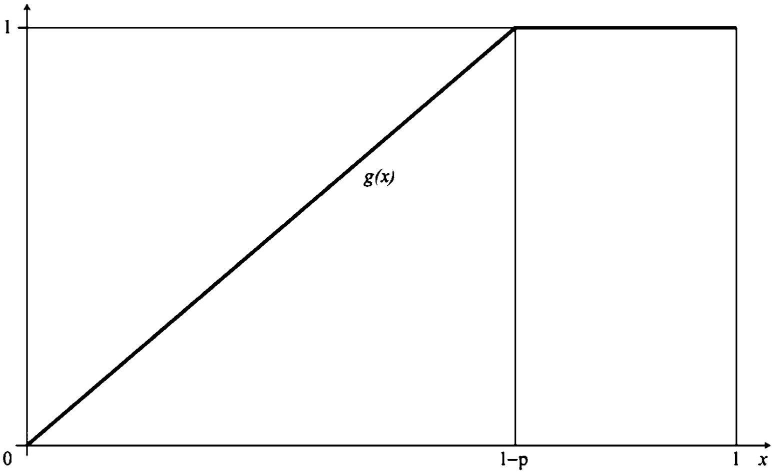


Fig. 25.2 Distortion function of CVaR

The corresponding risk measure WT_q is known as the *Wang transform*. The parameter q can be changed to make the Wang transform either sharper on high losses or softer and more receptive to positive returns. Wang (2002) recommended the Wang transform for the measurement of insurance risks.

- The beta family of distortion risk measures, proposed by Wirch and Hardy (1999), utilizes the incomplete beta function:

$$g(F_X(x)) = \beta(a,b;F_X(x)) = \int_0^{F_X(x)} \frac{1}{\beta(a,b)} t^{a-1}(1-t)^{b-1}\,\mathrm{d}t = S_\beta(F_X(x)),$$

where $S_\beta(x)$ is the distribution function of the beta distribution and $\beta(a,b)$ is the beta function with parameters $a > 0$ and $b > 0$; that is

$$\beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 t^{a-1}(1-t)^{b-1}\,\mathrm{d}t.$$

The beta-distortion risk measures are concave if and only if $a \leq 1$ and $b \geq 1$; strictly concave if a and b are both not equal to 1.

- The *Proportional Hazard* (PH) *transform* is a special case of the beta-distortion risk measure with $a = 0.1$, $b = 1$. The PH-transform risk measure is defined as:

$$\rho_{\mathrm{PH}}(X) = \int_0^\infty S_X(x)^{\frac{1}{\gamma}}\,\mathrm{d}x, \quad \gamma > 1,$$

where $S_X(x) = 1 - F_X(x)$.

25.6 New Distortion Risk Measures

The notion of asymmetry of the risk perception of investors was studied in classical works such as (Kahneman and Tversky (1979)). Here we apply the idea of asymmetry to the standard construction of the distortion risk measures. Introducing the asymmetry into the Choquet integral, we obtain the following distortion risk measures:

$$\rho_{g_i}(X) = -\int_{-\infty}^{0} [1 - g_1(S_X(x))]\,\mathrm{d}x + \int_{0}^{\infty} g_2(S_X(x))\,\mathrm{d}x,$$

where distortion functions g_1 and g_2 only differ by their risk-averse parameters.

Moreover, the risk-averse parameter should be inserted in the properly chosen distortion function. Here we propose the quadratic distortion function

$$g_i(x) = x + k_i(x - x^2),$$

where $k_i \in (0, 1]$ is the risk-averse or concave parameter. The investor is more risk-averse with k closer to one. The chosen quadratic distortion function $g\colon [0, 1] \to [0, 1]$ comes within all the criteria of a "good" distortion function: it is continuous, differentiable, and strictly concave[22] when $k \in (0, 1]$. A strictly concave function leads to consistency with respect to SSD. In the case when $k = 0$, the distortion function equals the mathematical expectation $g(x) = x$.

As a result, we obtain a new asymmetric distortion risk measure based on the quadratic distortion function:

$$\rho_{g_i}(X) = -\int_{-\infty}^{0} [1 - g_1(S_X(x))]\,\mathrm{d}x + \int_{0}^{\infty} g_2(S_X(x))\,\mathrm{d}x.$$

where $g_i(x) = x + k_i(x - x^2)$, $k_i \in (0, 1]$, and $i = 1, 2$, k_1, k_2 are changing independently. The proposed risk measure treats upside and downside risk differently. The motivation for the introduction of asymmetry is the importance for the risk-averse investor of having $k_1 > k_2$ in order to put more weight on the left tail (losses), than on the right (gains).

The power function is widely used in economic theory, that is why it seems to be promising to use this function in the proposed framework of asymmetric distortion risk measures. We will apply the following form of the power distortion function:

$$g = x^k, k \in (0, 1).$$

---

[22] $g''(x) = -2k < 0$ at all points.

Risk-averse investors using the power distortion function to describe their risk perception will choose k closer to 0. The asymmetric distortion risk measured with the power distortion function will take the form:

$$\rho_{g_i}(X) = -\int_{-\infty}^{0} [1 - g_1(S_X(x))]\, \mathrm{d}x + \int_{0}^{\infty} g_2(S_X(x))\, \mathrm{d}x,$$

where $g_i(x) = x_i^k$, $k_i \in (0, 1)$, $i = 1, 2$, k_1, k_2 are changing independently. We again include the asymmetry by introducing different parameters on the left and right sides of the integral.

25.7 Summary

The natural question that arises for asset managers is the choice of an adequate risk measure. The answer to this question is not obvious, as it is generally not easy to identify which particular risk measure might be the best and there is no clear way of comparing one risk measure to another. Furthermore, there is no guarantee that an arbitrarily chosen measure would necessarily be “good.”

In the paper, the class of distortion risk measures is analyzed. It possesses the most desirable properties for a portfolio risk measure: law-invariance, subadditivity, and consistency with the second-order stochastic dominance. In addition, distortion risk measures have their roots in the distortion utility theory of choice under uncertainty, meaning that this class of risk measures can better reflect the risk preferences of investors.

The well-known examples of distortion risk measures were reviewed and the drawbacks of VaR and CVaR in the capacity of distortion function were explained. Moreover, we introduce new asymmetric distortion risk measures that possess the property of asymmetry along with the standard properties of concave distortion risk measures. This new risk measures reflect the entire range of an investor’s preferences.

25.8 Appendix: Properties of Risk Measures

Axioms to characterize a particular risk measure are usually necessary to obtain mathematical proofs. They can generally be divided into three types (Denuit et al. 2006):

- *Basic rationality axioms* are satisfied by most of the risk measures (e.g., monotonicity)
- *Additivity axioms* include sums of risks (e.g., subadditivity, additivity, and superadditivity)
- *Technical axioms* deal mostly with continuity conditions

None of the following properties is absolute. Almost all of them are subject to criticism.

Property 25.1. Law-invariance
Law-invariance states that a risk measure $\rho(X)$ does not depend on a risk itself but only on its underlying distribution, i.e., $\rho(X) = \rho(F_X)$, where F_X is the distribution function of X. This condition ensures that F_X contains all the information needed to measure the riskiness of X. Law-invariance can be phrased as:

$$F_X = F_Y \Rightarrow \rho(X) = \rho(Y)$$

for every random portfolio returns X and Y with distribution functions F_X and F_Y. In other words, ρ is law invariant in the sense that $\rho(X) = \rho(Y)$, whenever X and Y have the same distribution with respect to the initial probability measure, P. This assumption is essential for a risk measure to be estimated from empirical data, which ensures its applicability in practice.

Property 25.2. Positive homogeneity
Positive homogeneity (also known as positive scalability) formulates as follows: for each positive λ and random portfolio return $X \in \mathcal{X}$:

$$\rho(\lambda X) = \lambda^k \rho(X).$$

Positive homogeneity signifies that a measure has the same dimension (scalability) as a variable X. When the parameter $k = 0$, a risk measure does not depend on the scalability.

From a financial perspective, positive homogeneity implies that a linear increase of the return by a positive factor leads to a linear increase in risk by the same factor. Although Artzner et al.. (1999) require adequate risk measures to satisfy the positive homogeneity property, Föllmer and Schied (2002a, b) drop this assumption arguing that risk may grow in a nonlinear way as the size of the position increases. This would be the case with liquidity risk. Dhaene et al. (2003) and De Giorgi (2005) also do not believe that this rule characterizes rational decision-makers' perception of risk.

Property 25.3. Sums of risks
Consider two different financial instruments with random payoffs $X, Y \in \mathcal{X}$. The payoff of a portfolio consisting of these two instruments will equal $X + Y$.

Property 25.3.1 Subadditivity
Subadditivity states that the risk of the portfolio is not greater than the sum of the risks of the portfolio components. In other words, "a merger does not create extra risk" (Artzner et al.. 1999).

$$\rho(X + Y) \leq \rho(X) + \rho(Y).$$

Compliance with this property tends to the diversification effect. Though Artzner et al. (1999) treat subadditivity as a necessary requirement for constructing a risk measure in order for it to be coherent, empirical evidence suggests that subadditivity does not always hold in reality.[23]

Property 25.3.2 Additivity
The additivity property is expressed in the following form:

$$\rho(X+Y)=\rho(X)+\rho(Y).$$

This property is valid for independent and comonotonic[24] random variables X and Y. The comonotonic random variables with no-hedge condition result in *comonotonic additivity*.

Property 25.3.3 Superadditivity
Superadditivity states that the portfolio risk estimate could be greater than the sum of the individual risk estimates.

$$\rho(X+Y)\geq\rho(X)+\rho(Y).$$

The superadditivity property is valid for risks which are positive (negative) dependent.

Property 25.4. Convexity

1. For all $X, Y \in \mathcal{X}$, $0 \leq \lambda \leq 1$, the following inequality is true:

$$\rho(\lambda X+(1-\lambda)Y)\leq\lambda\rho(X)+(1-\lambda)\rho(Y).$$

Convexity ensures the diversification property and relaxes the requirement that a risk measure must be more sensitive to aggregation of large risks.

2. For any $\lambda, \mu \geq 0$, $\lambda+\mu=1$, and distribution functions F, G, the following inequality holds

$$\rho(\lambda F+\mu G)\leq\lambda\rho(F)+\mu\rho(G).$$

3. *Generalized convexity*. For any $\lambda, \mu \geq 0$, $\lambda+\mu=1$, and distribution functions U, V, H, such that the following random variables exist $X, Y, \lambda X+\mu Y$, for which $F_X=U$, $F_Y=V$, $F_{\lambda X+\mu Y}=H$, the inequality is true

$$\rho(H)\leq\lambda\rho(U)+\mu\rho(V).$$

[23] Critiques of subadditivity can be found in Dhaene et al. (2003) and Heyde et al. (2006).

[24] Comonotonic or common monotonic random variables (Yaari 1987; Schmeidler 1986; Dhaene et al., b) are those such that if the increase of one follows the increase of the other variable:

$$P[X\leq x, Y\leq y]=\min\{P[X\leq x], P[Y\leq y]\} \quad \text{for all } x, y\in\mathbb{R}.$$

Intuitively, such variables have a maximal level of dependency. They are positively correlated so that the correlation coefficient approximates to one. In financial and insurance markets, this property appears quite frequently.

Property 25.5. Monotonicity
For every random portfolio returns X and Y such that $X \geq Y$,

$$\rho(X) \leq \rho(Y).$$

Monotonicity implies that if one financial instrument with the payoff X is not less than the payoff Y of the other instrument, then the risk of the first instrument is not greater than the risk of the second financial instrument. Another presentation of the monotonicity property with a risk-free instrument is as follows:

$$X \geq 0 \Rightarrow \rho(X) \leq \rho(0)$$

for $X \in \mathcal{X}$.

Property 25.6. Translation invariance

Property 25.6.1 For the nonnegative number $\alpha \geq 0$ and $C \in \mathbb{R}$, the property has the following form:

$$\rho(X + C) = \rho(X) - \alpha C.$$

This property states that if the payoff increases by a known constant, the risk correspondingly decreases. In practice, $\alpha = 0$ or $\alpha = 1$ are often used.

Property 25.6.2 When $\alpha = 0$, it implies that the addition of a certain wealth does not increase risk. This property is also known as the *Gaivoronsky–Pflug (G–P) translation invariance* (Gaivoronski and Pflug 2005).

Property 25.6.3 The case when $\alpha = 1$ implies that by adding a certain payoff, the risk decreases by the same amount.

$$\rho(X + C) = \rho(X) - C.$$

Property 25.6.4 When a constant wealth has a positive value, i.e., $C \geq 0$, one gets

$$\rho(X + C) \leq \rho(X).$$

This result is in agreement with the monotonicity property of $X + C \geq X$.

Property 25.6.5 In particular, translation invariance involves

$$\rho\left(X + \rho(X)\right) = \rho(X) - \rho(X) = 0,$$

obtaining a risk-neutral position by adding $\rho(X)$ to the initial position X.

Property 25.7. Consistency

Property 25.7.1 Consistency with respect to *n-order stochastic dominance* has the following general form:

$$X \geq_n Y, \rho(X) \geq \rho(Y).$$

In practice, the maximal value of $n = 2$; $n = 0$ just stands for a monotonicity property.

Property 25.7.2 Monotonic dominance of n-order

$$X \geq_{M(n)} Y, \text{ iff } E[u(X)] \geq E[u(Y)]$$

for any monotonic of order n functions, that is $u^{(n)}(t) \geq 0$. It is known that $X \geq_1 Y$ is equivalent to $X \leq_{M(1)} Y$. $X \leq_{M(2)} Y$ is also called the Bishop-de Leeuw ordering or Lorenz dominance.

Property 25.7.3 First-order stochastic dominance (FSD)

$$\text{For } X \geq_1 Y,\ F_X(x) \leq F_Y(x).$$

If an investor prefers X to Y, then FSD will indicate that the risk of X is less than the risk of Y. In terms of utility function u, the following holds:

$$\text{If } X \geq_1 Y, \text{ then } E[u(X)] \geq E[u(Y)]$$

for all increasing utility functions u. FSD characterizes the preferences of risk-loving investors. Ortobelli et al. (2008) classified risk measures consistent with respect to FSD as a *safety-risk measures.*[25]

Property 25.7.4 Rothschild–Stiglitz stochastic order dominance (RSD)
RSD was introduced by Rothschild and Stilglitz (1970) and has the form:

$$\text{If } X \leq_{\text{RS}} Y, \text{ then } E[u(X)] \geq E[u(Y)]$$

for any concave, not necessarily decreasing, utility function u. RSD describes preferences of risk-averse investors. *Dispersion measures* are normally consistent with RSD.

Property 25.7.5 Second-order stochastic dominance (SSD)
The concept of SSD was introduced by Hadar and Russell (1969), although Rothschild and Stilglitz (1970) first proposed its use in portfolio theory. SSD has the following form:

$$\text{For } X \geq_2 Y,\ E[u(X)] \geq E[u(Y)]$$

for all increasing, concave utility functions u. SSD characterizes nonsatiable risk-averse investors.

---

[25] In the portfolio selection literature, two disjoint categories of risk measures are defined: dispersion measures and safety-first risk measures. For the definitions and properties of specified categories, see for example, Giacometti and Ortobelli (2004).

Property 25.7.6 Stochastic order – stop-loss
Y dominates X ($Y \geq_{\text{SL}} X$) in stop-loss order, if for any number α the following inequality is true:

$$E[(Y-\alpha)^+] \geq E[(X-\alpha)^+].$$

Here $\alpha^+ = \max\{0, \alpha\}$. Such order is essential in the insurance industry. If the insurer takes the responsibility for the claims greater than α (deductible), then the expected claim Y is not smaller than X.

Property 25.7.7 Convex order
Y dominates X with respect to convex order ($Y \geq_{\text{CX}} X$), if the relation $Y \geq_{\text{SL}} X$ is true and when $\alpha = -\infty$ in stop-loss order, i.e., $E[X] = E[Y]$. Convex ordering is related to the notion of risk aversion.[26]

Consistency with the stochastic dominance is a necessary property for a risk measure because it enables one to characterize the set of all optimal portfolio choices when either wealth distributions or expected utility functions depend on a finite number of parameters (Ortobelli 2001).

Property 25.8. Nonnegativity

Property 25.8.1 $\rho(X) \geq 0$, while $\rho(X) > 0$ for all nonconstant risk.

Property 25.8.2 If $X \geq 0$, then $\rho(X) \leq 0$; if $X \leq 0$, then $\rho(X) \geq 0$.

Property 25.9. Continuity

Property 25.9.1 Probability convergence continuity: If $X_n \xrightarrow{\text{P}} X$, then $\rho(X_n)$ converges and has the limit $\rho(X)$.

Property 25.9.2 Weak topology continuity: If $F_X \xrightarrow{\text{w}} F_X$, then $\rho(F_{X_n})$ converges and has a limit $\rho(F_X)$.

Property 25.9.3 Horizontal shift continuity: $\lim_{\delta \to 0} \rho(X+\delta) = \rho(X)$.

Property 25.9.4 Opportunity of arbitrary risk approximation with the finite carrier is expressed by the equality[27]:

$$\lim_{\delta \to +\infty} \rho(\min\{X, \delta\}) = \lim_{\delta \to -\infty} \rho(\max\{X, \delta\}) = \rho(X).$$

Property 25.9.5 Lower semicontinuity: For any $C \in \mathbb{R}$, the set $\{X \in \mathcal{X}: \rho(X) \leq C\}$ is $\sigma(L^\infty, L^1)$ – closed.

---

[26] See also Kaas et al.. (1994, 2001).

[27] Wang et al. (1997) and Hürlimann (1994).

Property 25.9.6 Fatough property[28]

For any bounded sequence (X_n) for which $X_n \xrightarrow{\mathrm{P}} X$, the following holds:

$$\rho(X) \leq \lim_{n\to\infty} \inf \rho(X_n).$$

These properties are cardinally important. Non-fulfillment of the continuity property implies that even a small inaccuracy in a forecast can lead to the poor performance of a risk measure.

Property 25.10. Strictly expectation-boundedness
The risk of a portfolio is always greater than the negative of the expected portfolio return.

$$\rho(X) \geq -E[X], \text{ while } \rho(X) > -E[X] \text{ for all nonconstant } X,$$

where $E[X]$ is the mathematical expectation of X.

Property 25.11. Lower-range dominated
Deviation measures possess lower-range dominated property of the following form:

$$D(X) \leq E(X)$$

for a nonnegative random variable. From Properties 25.10 and 25.11 one can derive:

$$D(X) = \rho(X - EX), \quad \rho(X) = D(X) - E(X).$$

Property 25.12. Risk with risk-free return C

Property 25.12.1 $\rho(C) = -C$, it follows from the *Invariance Property 25.6.3*. If $C > 0$, then the situation is stable, risk is negative. The opposite situation occurs with $C < 0$.

Property 25.12.2 $\rho(C) = 0$, risk does not deviate with the zero certain return.

According to the classification given by Albrecht (2004), a number of risk measures can be divided into two categories – measures of the "first kind" and "second kind" – subject to the type of risk conception. Risk measures with Properties 25.2 and 25.12.2 belong to the "first kind," where risk is perceived as the quantity of deviations from a target. Risk measures with Properties 25.2 and 25.12.1 are of the "second kind," where risk is considered as a "necessary capital respectively necessary premium."

[28] In some contexts, it is equivalent to the upper semicontinuity condition with respect to $\sigma(L^\infty, L^1)$.

Property 25.13. Symmetric property

1. $\rho(-X) = -\rho(X)$, which corresponds to Property 25.8.1.
2. $\rho(-X) = \rho(X)$, this property makes sense for the measures with possible negative values (Property 25.8.2 fulfilled).

Property 25.14. Allocation
A risk measure need not be defined on the whole set of values of a random variable. Formally, in a given set U, from the condition $F_X = F_Y$, when $x \notin U$, it follows that $\rho(X) = \rho(Y)$. Apparently, this property holds only for law-invariant measures. Most often, some threshold value T is assigned, and the set U takes values $U = (-\infty, T]$ or $U = [T, \infty)$.

Property 25.15. Static and dynamic natures
It is useful to use a dynamic and multiperiod framework to answer the following question: How should an institution proceed with new information in each period and how should it reconsider the risk of the new position? Riedel (2004) introduced the specific axioms such as predictable translation invariance and dynamic consistency for a risk measure to capture the dynamic nature of financial markets.

References

Acerbi C., 2002, Spectral measures of risk: a coherent representation of subjective risk aversion, *Journal of Banking and Finance* 26, 1505–1518.

Acerbi C. and Tasche D., 2004, On the coherence of expected shortfall, *Journal of Banking and Finance* 26, 1487–1503.

Adam A., Houkari M. and Laurent J., 2007, Spectral Risk Measures and Portfolio Selection, http://hal.archives-ouvertes.fr/docs/00/16/56/41/PDF/Adam-Houkari-Laurent-ISFA-WP2037.pdf.

Albrecht P., 2004, Risk measures, in *Encyclopedia of Actuarial Science*, Wiley, New York.

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

Balbas A., Garrido J. and Mayoral S., 2007, Properties of Distortion Risk Measures, available at http://www.gloriamundi.org/picsresources/bgm_pdr.pdf.

Balzer L., 2001, Investment risk: a unified approach to upside and downside returns, in *Managing Downside Risk in Financial Markets*: *Theory, Practice and Implementation*, eds. Sortino F. and Satchell S., 103–105. Oxford: Butterworth-Heinemann.

Bäuerle N. and Müller A., 2006, Stochastic orders and risk measures: consistency and bounds, *Insurance: Mathematics and Economics* 38(1), 132–148.

Bellini F. and Caperdoni C., 2006, *Coherent distortion risk measures and higher order stochastic dominances*, preprint, January 2006, available at http://www.gloriamundi.org.

Chekhlov A., Uryasev S. and Zabarankin M., 2003, Portfolio optimization with Drawdown constraints, Implemented in Portfolio Safeguard by AORDA.com, *Theory Probability Application* 44(1).

De Giorgi E., 2005, Reward-risk portfolio selection and stochastic dominance, *Journal of Banking and Finance* 29(4), 895–926.

Denneberg D., 1990, Distorted probabilities and insurance premiums, *Methods of Operations Research* 63, 3–5.

Denneberg D., 1994, *Non-Additive Measures and Integrals*. Dordrecht: Kluwer.

Denuit M., Dhaene J., Goovaerts M. and Kaas R., 2005, *Actuarial Theory for Dependent Risks: Measures, Orders and Models*. Chichester: Wiley.

Denuit M., Dhaene J., Goovaerts M., Kaas R. and Laeven R., 2006, Risk measurement with equivalent utility principles, *Statistics and Decision* 24, 1–25.

Dhaene J. and Goovaerts M., 1996, Dependency of risks and stop-loss order, *Astin Bulletin* 26(2), 201–212.

Dhaene J., Denuit M., Goovaerts M., Kaas R. and Vyncke D., 2002a, The concept of comonotonicity in actuarial science and finance: theory, *Mathematics and Economics* 31, 3–33.

Dhaene J., Denuit M., Goovaerts M., Kaas R. and Vyncke D., 2002b, The concept of comonotonicity in actuarial science and finance: applications, *Mathematics and Economics* 31, 133–161.

Dhaene J., Goovaerts M. and Kaas R., 2003, Economic capital allocation derived from risk measures, *North American Actuarial Journal* 7(2), 44–59.

Ebert U., 2005, Measures of downside risk, *Economics Bulletin, Economics Bulletin* 4(16), 1–9.

Embrechts P., McNeal A. and Straumann D., 2002, Correlation and dependence in risk management: properties and pitfalls, in *Risk Management: Value at Risk and Beyond*, ed. Dempste M.A.H., 176–223. Cambridge: Cambridge University Press.

Fabozzi F. and Tunaru R., 2008, Pricing Models for Real Estate Derivatives, Yale Working paper.

Föllmer H. and Schied, A., 2002a, Convex measures of risk and trading constraints, *Finance and Stochastics* 6(4), 429–447.

Föllmer H. and Schied, A., 2002b, Robust preferences and convex measures of risk, in *Advances in Finance and Stochastics*, eds. Sandmann K. and Schonbucher P. J., 39–56. New York: Springer.

Frittelli M. and Rosazza Gianin E., 2005, Law invariant convex risk measures, *Advances in Finance and Stochastics* 7, 33–46.

Gaivoronski A. and Pflug G., 2005, Value at Risk in Portfolio Optimization: Properties and Computational Approach, *Journal of Risk*, 7, 1–31.

Giacometti R. and Ortobelli S., 2004, Risk measures for asset allocation models, in *Risk Measures for the 21st Century*, ed. Szegö G., 69–87. Chichester: Wiley.

Goovaerts M., De Vijlder F. and Haezendonck J., 1984, *Insurance Premiums*. Amsterdam: North-Holland.

Goovaerts M., Kaas R., Dhaene J. and Qihe T., 2003, A unified approach to generate risk measures, *Astin Bulletin* 33, 173–191.

Gourieroux C. and Liu W., 2006, *Efficient portfolio analysis using distortion risk measures*, Les Cahiers du CREF 06–35.

Hadar J. and Russell W., 1969, Rules for ordering uncertain prospects, *American Economic Review* 59, 25–34.

Hamada M., Sherris M. and van Der Hoek J., 2006, Dynamic portfolio allocation, the dual theory of choice and probability distortion function, *Astin Bulletin* 36(1), 187–217.

Hardy M. and Wirch J., 2003, Ordering of risk measures for capital adequacy, in Proceedings of the AFIR Colloquim, Tromsoe.

Heyde C., Kou S. and Peng X., 2006, What Is a Good Risk Measure: Bridging the Gaps Between Data, Coherent Risk Measure, and Insurance Risk Measure, Columbia University.

Hürlimann W., 2004, Distortion risk measures and economic capital, *North American Actuarial Journal* 8(1), 86–95.

Inoue A., 2003, On the worst conditional expectation, *Journal of Mathematical Analysis and Applications* 286, 237–247.

Kaas R., Van Heerwaarden A. and Goovaerts M., 1994, *Ordering of actuarial risks*, CAIRE Educations Series, Brussels.

Kaas R., Goovaerts M., Dhaene J. and Denuit M., 2001, *Modern actuarial risk theory*. Dordrecht: Kluwer.

Kahneman D. and Tversky A., 1979, Prospect theory: An analysis of decision under risk, *Econometrica 47*, 313–327.

Kusuoka E., 2001, *Law invariant risk measures have the Fatou property*, preprint, Universite Paris Dauphine.

Markowitz H., 1952, Portfolio selection, *Journal of Finance*, 7(1), 77–91.

Ortobelli S., 2001, The classification of parametric choices under uncertainty: analysis of the portfolio choice problem, *Theory and Decision* 51, 297–327.

Ortobelli S., Rachev S., Shalit H. and Fabozzi F., 2008, Risk Probability Functionals and Probability Metrics in Portfolio Theory, *Probability and Mathematical Statistics*, 28, 203–234.

Pedersen C. and Satchell S., 1998, An extended family of financial-risk measures, *The Geneva Papers on Risk and Insurance Theory* 23, 89–117.

Rachev R., Ortobelli S., Stoyanov S., Fabozzi F. and Biglova A., 2008, Desirable properties of an ideal risk measure in portfolio theory, *International Journal of Theoretical and Applied Finance.*

Riedel F., 2004, Dynamic coherent risk measures, *Stochastic Processes and their Applications*, 112, 185–200.

Rockafellar R., Uryasev S. and Zabarankin M., 2002, Deviation measures in generalized linear regression, *Risk Management and Financial Lab*.

Rockafellar R., Uryasev S. and Zabarankin M., 2003, Deviation measures in risk analysis and optimization, *Risk Management and Financial Engineering Lab*.

Rockafellar R., Uryasev S. and Zabarankin M., 2006, Generalized deviations in risk analysis, *Finance and Stochastics* 10(1), 51–74.

Rothschild M. and Stiglitz J., 1970, Increasing risk. I.a. definition, *Journal of Economic Theory* 2, 225–243.

Schmeidler D., 1986, Subjective probability and expected utility without additivity, *Econometrica* 57(3), 571–587.

Stone B., 1973, A general class of three-parameter risk measures, *Journal of Finance* 28, 675–685.

Szegö G., 2004, On the (non)acceptance of innovations, in *Risk Measures for the 21st Century*, ed., Szegö G., 1–10. Chichester: Wiley.

Tsanakas A. and Desli E., 2003, Risk measures and theories of choice, *British Actuarial Journal* 9, 959–981.

Van der Hoek J. and Sherris M., 2001, A class of non-expected utility risk measures and implications for asset allocation, *Insurance: Mathematics and Economics* 28(1), 69–82.

Wang S., 2002, *A risk measure that goes beyond coherence*, 12th AFIR International Colloquium, Mexico.

Wang S., 2004, *Cat bond pricing using probability transforms*, Geneva Papers, Special issue on Insurance and the State of the Art in Cat Bond Pricing 278, 19–29.

Wang S., Young V. and Panjer H., 1997, Axiomatic characterization of insurance prices, *Insurance: Mathematics and Economics* 21, 173–183.

Wirch J. and Hardy M., 1999, A synthesis of risk measures for capital adequacy, *Insurance: Mathematics and Economics* 25, 337–347.

Yaari M., 1987, The dual theory of choice under risk, *Econometrica* 55, 95–116.

Zhang Y. and Rachev S., 2006, Risk Attribution and Portfolio Performance Measurement: An Overview, *Journal of Applied Functional Analysis*, 4, 373–402.

Chapter 26
A Benefit from the Modern Portfolio Theory for Japanese Pension Investment

Makoto Suzuki

26.1 Introduction

In 1990, Dr. Harry Markowitz was awarded the Nobel Prize in Economics. His contribution was not only finding the new theory in the investment but also giving benefit to the real world through actual investment activities. Harry Markowitz developed the efficient frontier in his seminal 1952 article in which return is maximized relative to risk, as defined by the standard deviation. William Sharpe expanded on the Markowitz risk–return analysis by braking risk in to market risk and nonmarket risk. After CAPM was introduced, many investors focused on its implication, that individual stock returns were expressed as a function of the security by correlation coefficient (beta) with the market portfolio. Wells Fargo, which has been an asset management company, had a foresight to apply this result into the real investment. In 1971, Wells Fargo lunched the first index fund in the world.[1] The first investor of this index fund was the Samsonite pension fund, which invested six million dollars. It suggested that US pension funds would be a major investor to the index fund in the future. The way to replicate the market portfolio was equal weight investment on NYSE-listed stocks. In 1973, another idea for market portfolio replication was introduced. It was S&P500 stock index replicated portfolio. The S&P500 stock index was composed of 500 representative stocks in US stock market from 1957. Indexing would not be able to complete replication for the market portfolio; however, indexing made easier to set up and maintain the actual fund. Other pension funds which were Wells Fargo and Illinois Bell started to invest on S&P500 stock index truck fund.[2]

The index fund started with six million dollars grows more than one trillion dollars in equity and 215.5 billion dollars in fixed income securities at top 200 US

M. Suzuki (✉)
Research, Daiwa Institute of Research Ltd., Grand Tokyo North Tower 29th fl, 1-9-1, Marunouchi, Chiyodaku, 100-6756 Japan
e-mail: makoto.suzuki@rc.dir.co.jp

[1] Peter L. Bernstein, CAPITAL IDEAS, p. 247.

[2] Peter L. Bernstein, CAPITAL IDEAS, p. 247.

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_26,
© Springer Science+Business Media, LLC 2010

pension funds in 2007.[3] It is estimated that more than 1.5 trillion dollars for all US pension funds are invested in the index fund. One of a main reason for the growth amount of the index funds is a development of the modern finance theory. However, we need to recognize a role of Employees Retirement Income Security Act (ERISA) of 1974 to support the index fund investment, too. The goal of ERISA has been to secure a recipient's right and benefit on their pension. Before ERISA was introduced, pension fund was managed at employer's discretion. In ERISA, investment manager of the pension fund requested to be disciplined to follow the "prudent man rules." A lot of pension funds, not only private pensions but public pensions, weighed their investment to the index funds since the index fund has rationality and objectivity which were requested by "prudent man rule."

26.1.1 Contribution to Japan

The first index fund was launched by Kokusai asset management in 1985.[4] The majority of the investor was retail investor since Japanese pension funds were strictly regulated to invest in the risk assets in 1980s. There were some challenges for the Japanese pension funds to make their own investment policy by themselves. There were some conditions required to change the paradigm, first was a recognition improvement of modern finance theory and investment, second was a deregulation for risky assets investment and third was an introduction of fiduciary duty and prudent man rule. Usually a recognition improvement will take time; however, it was a good timing that Harry Markowitz, William Sharpe, and Merton Miller were awarded Nobel Prize in 1990. For Japanese, it was easy to know their name and achievements, after that. Further more, some Japanese financial institution started to operate their research base for new investment in US from late 1980s. Some Nobel Prize recipients and famous Professors at Finance were hired and studied at the institutions.[5] Finally, the modern finance theory and an investment vehicle, Index Fund, were well-known in Japan. The second point is a deregulation. The Pension Fund Association has started in 1966 and wholly entrusted to the trust banks and mutual life insurance companies on their investment by the government regulation. The asset allocation of those institutions had been regulated by the Ministry of Finance. The fund management at the trust banks had to follow the rules by the Ministry of Finance (Table 26.1). They had to allocate more than 50% to the risk-free assets, less than 30% for equity and less than 20% for real estate. On the mutual life insurance corporations, their investments were also regulated by the Insurance Business Procedure Rule such as fixed income securities investment more than 50%, equity

[3] John Bogle, 2000 "On investing: the first 50 years" McGraw-Hill and Pension and Investment, January 21, 2008.

[4] According to the web cite of the Asset Management Association of Japan.

[5] This author has an experience to study with Harry Markowitz for Japanese equity research in New Jersey in early 1990s.

Table 26.1 Investment rule for trust banks before deregulation

Risk-free asset (government bond, municipal bond, etc.)	More than 50%
Equity	Less than 30%
Foreign asset (equity and fixed income securities)	Less than 30%
Rear estate	Less than 20%

Table 26.2 The guide book of fiduciary duty (extraction) by the PFA

"There are four important elements, carefulness, faithfulness, self-execution, and separate management in fiduciary duty. Especially, carefulness is the same concept as a "prudent man rules" in United States and United Kingdom

investment less than 30%, and real estate investment less than 20%. Especially, since foreign exchange was strictly under controlled by Japanese government at that time, it was practically difficult to allocate to the foreign assets. In 1979, it was admitted for the mutual life insurance companies to allocate 10% for foreign assets. The foreign asset allocation had gradually expanded until 30% in 7 years. After the deregulation on pension fund management in 1997, the pension fund had more discretion on their investment.

Third, the fiduciary duty was introduced as a liability on pension fund managers. Under the strictly regulated asset allocation regime, it has diluted the feeling of liability on pension fund managers. However after deregulation, Ministry of Health, Labor and Welfare introduced "A guide line for fiduciary duty" in 1997 and the Pension Fund Association (PFA) distributed "The Handbook for Fiduciary duty." These efforts made to understand an importance of fiduciary duty on Japanese Pension managers. Even though there are no "ERISA"-type regulations in Japan, it will be applied by Common Law or other related Government orders for pension fund managers (Table 26.2).

Only when those three necessary conditions for investment environment are met, the Japanese pension funds would be able to invest at their discretion with own liability. Their investment for risky assets class, especially, Japanese equity and foreign equity, grew sharply over 40% after 1997 (Fig. 26.1).

26.1.2 Benefits for Japanese Pension Fund Investment

How we shall estimate the benefit from the Modern Finance Theory and deregulation? We could compare the asset allocation under regulation in 1996 and deregulated after 1997. Since the defined contribution (DC) type of pension system has been introduced after 2001, some portion of the defined benefit plan shifted their assets to DC plan; therefore, the total asset amount data does not continue. We assume that the plan asset in 1997 would continue to 2006. It is estimated that the

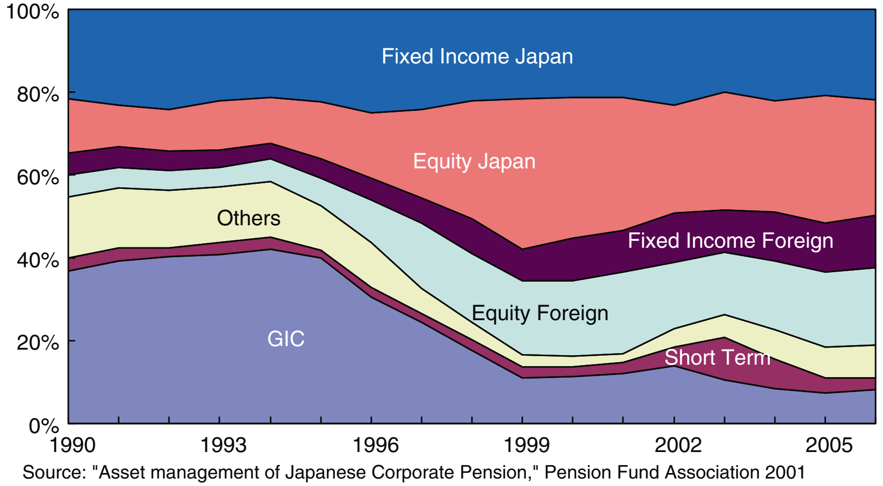


Fig. 26.1 Asset allocation on Japanese Private Pension Fund

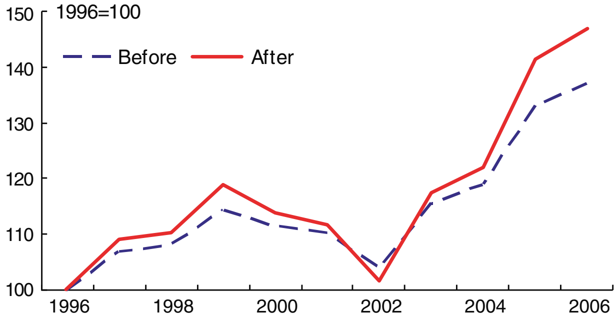


Fig. 26.2 Accumulated return (indexed)

difference between the regulated asset and deregulated asset was 4.2 trillion yen, 39.3 billion dollars, and about 10% in accumulated return basis (Fig. 26.2).[6]

26.1.3 Efficiency of the Market Model in Japanese Market

We have discussed about the necessary conditions, which are the modern theory recognition, the deregulation and weigh a liability on the manager. Now we need to confirm the sufficient condition side for the index portfolio investment.

[6] According to the PFA annual paper, estimated average annual returns are 3.2% (regulated asset allocation) and 3.9% (after deregulation).

We estimated the extra return on Japanese pension to about 39.3 billion dollars. However we should examine the superiority on index investment with more sophisticated method. Then, we focus on the examination of the market model efficiency in Japanese market between November1988 and November 2000.

26.1.3.1 Data

Our datasets for this examination are basically taken from Tokyo Stock Exchange listed companies, excluding financial and insurance, stock price and financial data amounts to a total of 394,815 records. We use shareholders equity, net earnings, and market capitalization as financial data and a yield of Certified Deposit for 3 months is applied to risk-free rate. Furthermore, we divide our examination period with Japanese economic cycle I–V.

The first period is between November 1988 and February 1991. It is officially recognized as an economic expansion period under the bubble and its adjustment. The second period is a contraction period between February 1991 and October 1993. From October 1993 through May 1997, the third period is a slow expansion period. During this period, there are many counter plans to meet the economic situation. In the fourth period, between May 1997 and January 1999, since a sales tax raised from 3 to 5%, an economic cycle turns to contraction. Final period is fifth period. An economic expansion[7] which is defined by the Economic and Social Research Institute of the Cabinet office continues until January 2002; however, we examined only till November 2000 because of data constraints (Table 26.3).

26.1.3.2 The Model

Theoretically, the background of the Index fund investment requires a condition which satisfies the CAPM. However, we will not further discuss regarding the CAPM and real market. We focus on analyzing a relation between rate of return on individual securities and market portfolio in Japanese market. According to the former study of Fama and French, the following three factors are significant on the return: book-to-market (B/M), earnings-to-price (EP), and market capitalization

Table 26.3 Dataset and economic cycle

Period	(I)	(II)	(III)	(IV)	(V)
Economic cycle	Expansion	Contraction	Expansion	Contraction	Expansion
Data set no.	11	12	13	14	15
	21		22		–
	3				

[7] Japanese economy gradually expanded under disinflation circumstance in fifth period.

(M, proxy for the scale) in the formula. We estimate the efficiency of the market portfolio by economic cycle with five regression models. Our basic market model with the three factors is as follows:

$$R_{\mathrm{i}} - R_{\mathrm{f}} = \alpha_{\mathrm{i}} + \beta_{\mathrm{i}}(R_{\mathrm{m}} - R_{\mathrm{f}}) + \gamma_{\mathrm{i}}\left(\frac{B_{\mathrm{i}}}{M_{\mathrm{i}}}\right) + \delta_{\mathrm{i}}\left(\frac{E_{\mathrm{i}}}{M_{\mathrm{i}}}\right) + \lambda_{\mathrm{i}}(\ln\ M_{\mathrm{i}}) + \varepsilon,$$

where

R: return of individual securities (i), R_{f}: risk free rate, R_{m}: return of market portfolio, B: shareholders' equity of security (i), M: market capitalization (i), E: net earnings (i), ε: Residual excess return.

26.1.3.3 Result of Estimation

Model 1: Estimate a risk premium by the market model

We employ the following model, which is transformed from the original market model, $R_{\mathrm{i}} - R_{\mathrm{f}} = \alpha_{\mathrm{i}} + \beta_{\mathrm{i}}(R_{\mathrm{m}} - R_{\mathrm{f}}) + \varepsilon$ for the first estimation. According to the results of estimation, β[8] which shows a slope of risk premium are statistically significant and positive all over the periods. A minimum beta is 0.81 in period V and a maximum is 1.28 in period IV. We estimated a value of 1.06 for beta when we combined all the dataset, which was data set no. 3. It is also statistically significant at one percentile point. The R-square are between 0.22 and 0.42, excluding period V. These results are coincident with the empirical research on Japanese market. Therefore, we could say at least between excess return and risk premium of the market portfolio we have stable positive correlation, since the estimation results are statistically significant. It suggests that the market model would be acceptable under our estimation.

Model 2: The market model with book (shareholder's equity)-to-market ratio

In the model 2, we added a book-to-market (B/M) ratio in the original model. We assume that a market risk premium and B/M ratio have liner relation with excess return under model 2. According to our estimation, β is positive and statistically significant on every period. These results are same and supportive to efficiency of the market model. However γ which is an estimated coefficient of B/M shows negative signs in the period I, II and V but positive signs in period III and IV. Even though those estimated signs are different, they are still statistically significant. Therefore, we could not say the B/M ratio is a stable factor in the model. We are wondering why the sign changes in each period? We suppose a book value basis accounting system and a cross share holding are a part of the reason. The book value is reflected the legacy value not the "market value." On the other hand, the market capitalization links to the "market." The B/M ratio would depreciate under the stock price hike, and the B/M ratio would appreciate under the weak market. The movement of the

[8] Beta is calculated based on the market portfolio. The estimation periods are term I–V.

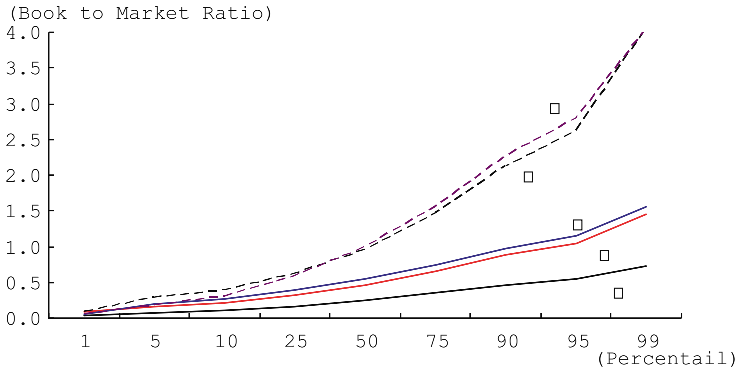


Fig. 26.3 Accumulated frequency on B/M ratio

B/M ratio seemed well influenced by the market. Figure 26.3 is an accumulated frequency on the B/M ratio for period I–V. We can see the line shifted from period I to the upward in the period II, III, IV and V gradually. Those shifts made some influence on our estimation. We shall discuss the B/M ratio puzzle later.

Model 3: The market model with earnings to price (EP) ratio

In this model, we add EP ratio instead of B/M ratio in the model 2. According to our estimation, signs of coefficient δ are different on each period. A positive signs are in the period II, III and V, negative signs are in the period I and IV. The coefficient δ is statistically significant but R-square does not improve from the model 1 on each period. On the other hand, β, the coefficient of market, is statistically significant and a stable positive sign. It would be difficult to say the EP ratio is an explanation variable for risk premium under those periods.

Model 4: The market model with market capitalization (M)

The market capitalization is a proxy for the company scale. Our estimation shows that the proxy of the scale works better than B/M and EP, because the estimated coefficient λ is statistically significant and its sign is positive in each period. Furthermore the coefficient of the market, β, is positive and significant under every period. Therefore, under our estimation we could say that the scale factor influences the excess return, and large-scale company stock would request more premium than small-scale company.[9]

Model 5: The market model with three factors

Finally, we examined multiregression with all factors. We found three results from this estimation. First, β, which is a coefficient of the market portfolio, seems stable and positive and statistically significant under all estimated periods. Second,

[9] Historically, the small scale company stock shows higher returns but the periods which we estimated are unusual economic condition, “deflation economy.” Most of the investors moved “flight to safety” in their investment and required additional premium on their investment, which are discussed in “Heisei Bable no Kenkyu (2002).”

the market capitalization, which is a proxy for the company scale, is statistically positive and stable in all periods. Third, the book-to-market ratio (B/M) and the earnings-to-price ratio (EP) show unstable sign on their coefficients, and is even statistically significant. Especially, since B/M and EP have multicollinearity relation under the period III, the sign shows different from the result of model 2.

According to the estimation, we shall conclude the results. First, the market portfolio coefficient β shows positive sign and statistically significant in all estimated periods. It shows that β is a most influential factor for risk premium in the Japanese equity market. Second, the company scales also have influence on the individual excess return. Third, the shareholder's equity (B/M) and net earnings (EP) do not show stable sign under the estimated period. It seems that those factors are difficult to place as the explanation variables for excess return in Japan.

Our conclusion raises an obvious question for inefficiency on the B/M ratio. The B/M ratio is relatively more stable factor to the EP and it is said that B/M explained excess return in Japan; however, it lose its power after 1990s. We shall discuss a possible reason for losing power on the B/M in the next.

26.1.3.4 Loosing Power of the B/M Ratio and Possible Explanation

The B/M ratio is calculated as the shareholder's equity divide by the market capitalization. A historical book value basis had been employed until April 2000 in Japan. Furthermore a cross-shareholdings were very common between affiliate companies and financial institutions. Therefore, some difference would be raised between market value and book value under the equity market condition. Since Japanese equity price grew sharply and born huge hidden profit on their cross-shareholdings until end of 1980s, it had been observed the "B/M ratio anomaly." After the bubble collapsed, the situation had completely turned around. The cross-shareholdings made a lot of implicit loss. However the book value changed inelastic but the market capitalization changed so fast. Therefore, the B/M ratio appreciated ($B/M > 1$) under such condition. Those conceptual ideas are on Figs. 26.4 and 26.5.

We examine the B/M ratio efficiency in the market model between November 1988 and April 2000. The collapse of the bubble period is included in the sample. According to the Nikkei[10] stock price chart in Fig. 26.5 (above), the stock market looked relatively stabilized after the bubble. However, even in the macro economic basis, implicit and explicit loss in the equity investment had been continued until end of our dataset (April 2000), according to the Systems of National Accounts. It would be difficult to say the write-off on cross-shareholding stocks had easily finished. The cross-shareholding amounts are huge in the large-scale companies than small-scale companies. Therefore, we need to segregate the companies from its scale. Historically, Japanese manufacturing companies organize a

[10] The Nikkei 225 Stock Price Index is one of the most popular stock indexes reported by Nikkei shin bun in Japan.

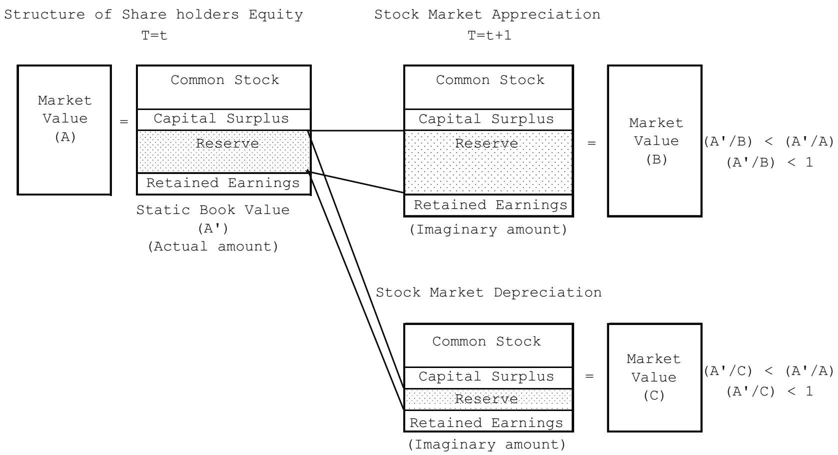
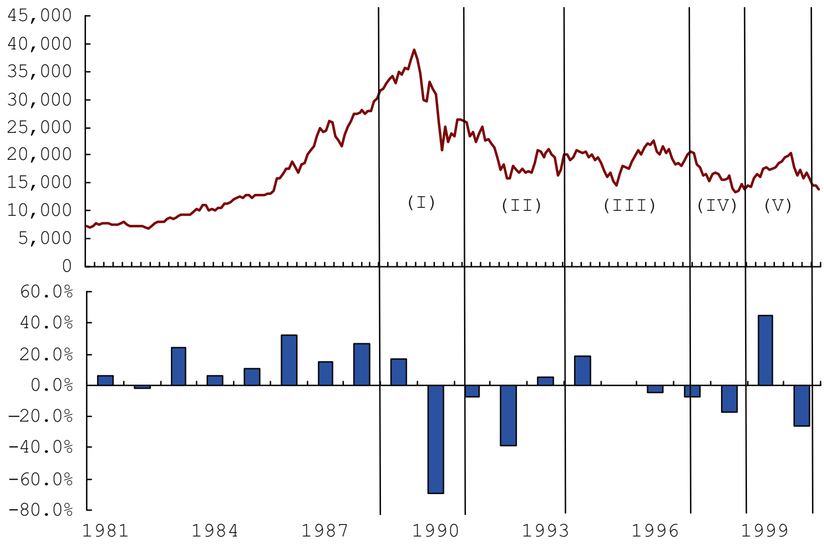


Fig. 26.4 The conceptual ideas for B/M ratio under market appreciation/depreciation



Source: Nikkei 225 Stock Price Index and Systems of National Accounts

Fig. 26.5 Stock price (*upper*) and profit and loss of the equity (lower). Source: Nikkei 225 stock price index and Systems of National Accounts

"Keiretsu[11]" to keep long-term relation with affiliate companies and to protect the merger and acquisition from an alien. Then we split our sample data by the median

---

[11] A Keiretsu is a kind of "Zaibatsu," "Konzern" or Holdings. It organized membership group and sometimes traded with mutually related companies.

point of the scale proxy in every period. Under the scale classification, we expect the B/M ratio shows more stable coefficient sign and statistically significant in the estimation.

According to an additional estimation by classification, the estimated coefficients of β and γ are stable and statistically significant in the large-scale group on every period. On the other hand, the results of the small-scale group are almost same as Model 2 Table 26.4. Therefore, it suggests that the result of Model 2 (Table 26.4) would be biased by the small-scale group. The low PBR (price-to-book value ratio) stock are placed as "value stock." However in our estimation, the B/M which is invert in relation with PBR shows negative coefficient sign in the large-scale group. This result is different from other former studies. It will be possible to understand that after the bubble collapsed, the large-scale companies had huge implicit loss in their share holder's book; therefore, the market discounts the premium. In other words, even though the stock market reflects the information, the book value on the financial account would change only every half year. On the other hand, the small-size group, they have limited cross-shareholdings compared to the large-scale group. Their implicit or explicit loss on cross-shareholdings is limited. Therefore, they have positive coefficient sign under the periods III and IV. In the periods III and IV, the implicit gain and loss are balanced on a macro basis (Fig. 26.5).

Furthermore, it will be drawn on the quadrant with total asset (Quadrant I), market capitalization (Quadrant II), shareholder's equity (Quadrant III), and B/M ratio (Quadrant IV) in Fig. 26.6. In this figure, Japanese companies are divided into two groups, large and small. It shows big difference between small-scale group and large-scale group in quadrant III. The line of B/M ratio is less than 1 in the small-scale group but another line crosses the line of the B/M ratio equal to 1. We suppose that share holder's equity moves inelastic than the market value; therefore, the B/M ratio is influenced by the price in the market and shifts upward. This circumstance will be explained that even though the book value of shareholder's equity does not move, the market capitalization (price) behavior makes the B/M ratio increase. Now, we will be able to understand from Table 26.5 that there is an inverse relation between movement of return and the B/M ratio with large-scale companies.

Next, we make a time series sample model in Fig. 26.7 for understanding this situation historically. Ideally, under an efficient market hypothesis, information will be reflected to the price quickly. However, the book value will be restated every half year. It is obvious that there is a time lag between market capitalization and book value. In Fig. 26.7, we put completely the same amount of market capitalization and shareholder's equity with time lag.

In Fig. 26.8, we made a model for explaining the difference between the ordinary B/M ratio and the ratio in 1990s. The trends of the B/M ratio shown in Fig. 26.8 seem same; however, the market capitalization and the shareholder's equity placed are different. An upper graph shows the ordinary case. When the B/M ratio shows the big difference between the market capitalization and the shareholder's equity, the market capitalization will be shifted by equity price appreciation. Then the B/M ratio reverts to the mean, which equals to one. On the other hand, lower graph shows the B/M ratio in 1990s under deflation economy. Investors expect that

Table 26.4 Results of the estimation

Model 1

Period	I	II	III	IV	V	I · II	III · IV	I–V
α	3.31	0.27	−0.11	−1.42	−2.16	1.63	−0.62	−0.33
t-value	40.66	4.42	−2.18	−13.07	−7.85	33.14	−13.05	−9.36
β	0.97	1.18	1.20	1.28	0.81	1.06	1.24	1.06
t-value	155.81	216.24	207.80	114.08	32.99	257.23	242.76	316.30
R^2	0.33	0.42	0.33	0.22	0.03	0.37	0.30	0.26

Model 2

Period	I	II	III	IV	V	I · II	III · IV	I–V
α	11.44	3.02	−0.11	−1.46	−1.95	5.41	−0.60	−0.34
t-value	73.58	28.98	−2.20	−13.38	−7.09	68.38	−13.16	−9.48
β	0.92	1.18	1.20	1.29	0.81	1.06	1.25	1.06
t-value	152.30	216.86	207.94	114.15	32.97	260.49	242.87	316.35
γ	−30.67	−5.35	0.01	0.03	−0.14	−9.23	0.01	0.01
t-value	−60.55	−32.29	12.40	4.47	−10.23	−60.43	11.07	7.15
R^2	0.38	0.43	0.33	0.22	0.03	0.39	0.30	0.26

Model 3

Period	I	II	III	IV	V	I · II	III · IV	I–V
α	4.82	0.24	−0.10	−1.42	−2.10	1.61	−0.64	−0.33
t-value	46.33	3.91	−2.04	−13.12	−7.65	32.67	−12.96	−9.28
β	0.93	1.18	1.20	1.28	0.81	1.06	1.24	1.06
t-value	154.96	216.39	207.97	114.08	33.02	257.28	242.85	316.40
δ	−79.24	1.25	0.01	−0.02	0.20	0.97	0.01	0.01
t-value	−23.00	7.12	13.79	−3.04	11.03	5.14	10.68	11.56
R^2	0.34	0.42	0.33	0.22	0.03	0.37	0.30	0.26

Model 4

Period	I	II	III	IV	V	I · II	III · IV	I–V
α	−0.92	−8.08	−8.21	−17.30	−24.33	−6.93	−12.77	−15.39
t-value	−1.45	−17.92	−22.47	−27.22	−25.52	−18.60	−39.94	−63.09
β	0.97	1.18	1.20	1.27	0.80	13.06	1.23	1.07
t-value	155.41	216.42	207.80	113.30	32.80	257.70	241.56	319.84
λ	0.37	0.77	0.76	1.56	2.20	0.77	1.15	1.41
t-value	6.70	18.68	22.38	25.36	24.27	23.19	38.43	62.38
R^2	0.33	0.42	0.33	0.23	0.05	0.37	0.31	0.27

Model 5

Period	I	II	III	IV	V	I · II	III · IV	I–V
α	10.25	−0.22	−7.48	−18.68	−23.57	3.27	−12.92	−15.05
t-value	15.93	−0.46	−20.48	−29.18	−24.64	8.29	−40.17	−61.50
β	0.92	1.17	1.20	1.28	0.80	1.06	1.24	1.07
t-value	152.22	217.90	208.34	114.12	32.81	262.19	241.77	320.03
γ	−30.03	−7.86	−0.25	0.20	−0.05	−11.84	0.06	−0.06
t-value	−55.50	−39.75	−18.97	15.17	−2.65	−67.71	7.25	−12.92
δ	−10.40	5.56	0.26	−0.19	0.13	7.56	−0.05	0.07
t-value	−2.92	27.07	19.80	−14.76	5.20	36.10	−6.03	14.93
λ	0.11	0.41	0.70	1.67	2.13	0.28	1.16	1.38
t-value	2.00	9.86	20.84	27.02	23.51	8.29	38.61	61.05
R^2	0.38	0.44	0.34	0.24	0.05	0.39	0.31	0.27

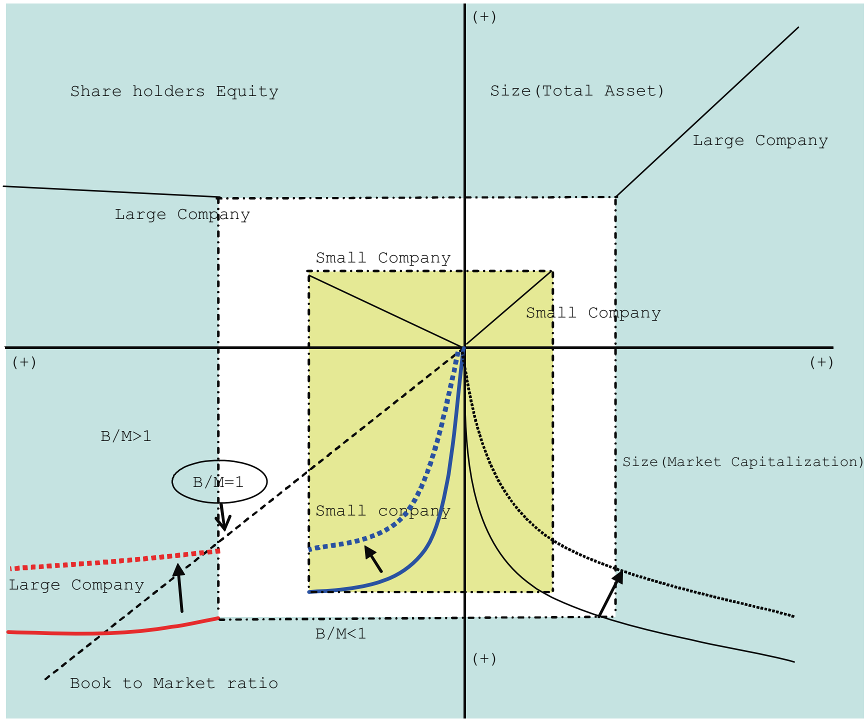


Fig. 26.6 Relation with total asset, shareholder's equity, market capitalization, and B/M ratio

Table 26.5 Result of estimation by the scale classification

Large market cap								
Period	I	II	III	IV	V	I · II	III · IV	I–V
α	11.54	5.15	6.32	4.94	12.80	7.97	5.61	7.40
t-value	50.62	29.65	42.24	24.07	223.05	65.72	47.80	89.65
β	0.95	1.07	1.14	1.08	0.76	1.02	1.09	1.04
t-value	111.55	159.48	161.19	88.01	20.32	199.52	178.88	253.08
γ	−29.14	−8.88	−10.42	−5.62	−13.50	−14.88	−8.32	−11.99
t-value	−36.86	−26.15	−40.92	−28.80	−32.24	−51.79	−47.02	−85.64
R^2	0.39	0.45	0.40	0.28	0.08	0.40	0.38	0.33
Small market cap								
Period	I	II	III	IV	V	I · II	III · IV	I–V
α	11.11	1.92	−0.91	−2.90	−5.04	3.68	−1.59	−2.16
t-value	51.23	12.88	−11.62	−15.94	−14.98	31.10	−22.22	−39.23
β	0.89	1.28	1.23	1.44	0.80	1.01	1.33	1.10
t-value	104.01	150.68	136.88	77.32	26.79	169.08	173.45	210.48
γ	−31.25	−4.22	0.01	0.04	−0.11	−6.85	0.01	0.01
t-value	−46.80	−20.65	11.10	4.87	−9.39	−35.92	9.86	7.15
R^2	0.36	0.42	0.30	0.21	0.04	0.37	0.28	0.24

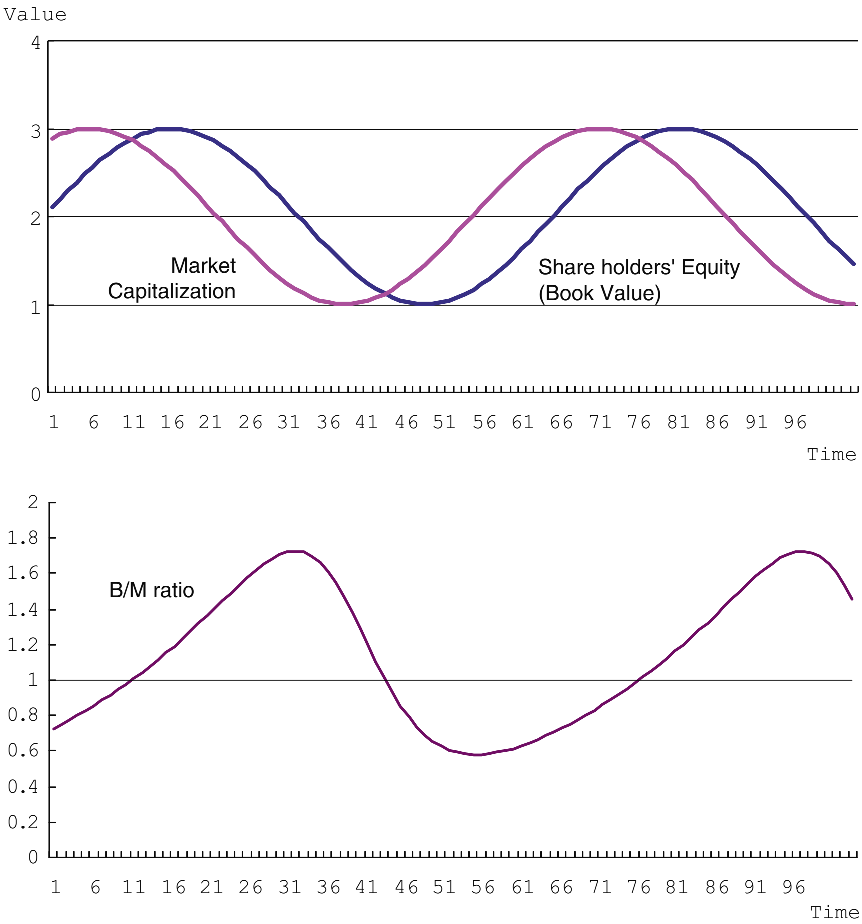


Fig. 26.7 Ideally model

some other implicit losses by cross-shareholdings will bail out and will influence the shareholder's equity. After restating the shareholder's equity, the B/M ratio decreases and reverts to its average. Under this situation, even though higher B/M ratio (lower price-to-book value ratio: PBR) will not add any extra return. This lower case suits to explain our results of estimation. The B/M ratio of large-scale company group is over one but estimated correlation coefficients are negative. This might be the reason for investors' expectation of implicit loss on the shareholder's equity. Since, large companies have relatively much cross-shareholdings than the small companies, investors suppose that they might have huge amount of hidden loss in their balance sheet. Investors rationally expect that the stock price will decline after the shareholder's equity is restated. Therefore, it seems a quite reasonable understanding for negative correlation coefficient to be statistically significant in large scale companies.

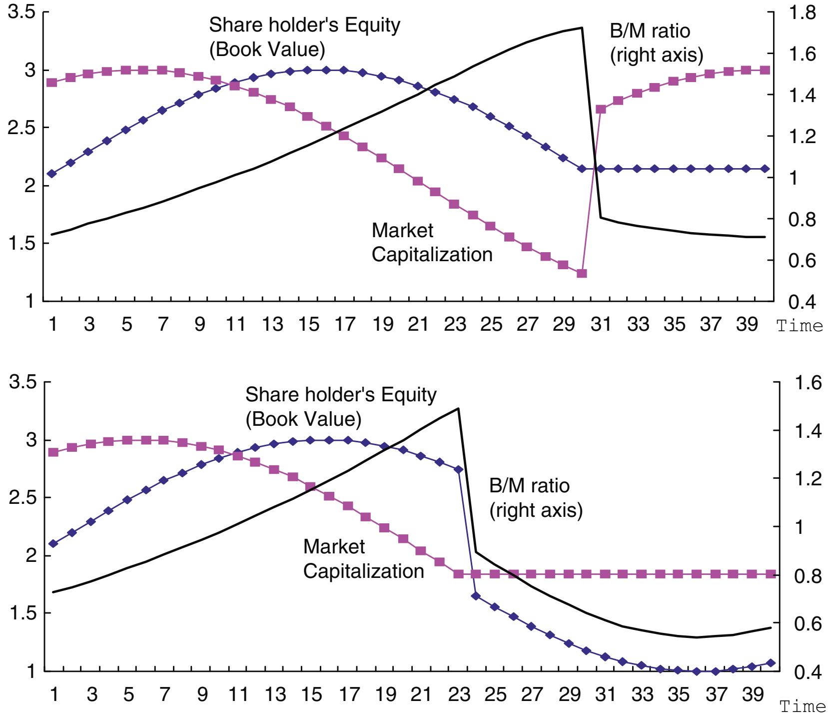


Fig. 26.8 Imaginary adjustment of B/M ratio

It could be concluded that the adjustment speed of an implicit gain and loss on the book value influences the sign of coefficient of B/M in the estimation. Furthermore, after April 2000, the Japanese accounting standard was changed to mark-to-market basis on holding securities. The hidden wealth of the company will not be able to store. Therefore, it is expected that the B/M ratio will no longer influence much on the excess return.

26.1.4 Conclusions

It is obvious that Harry Markowitz introduced a new paradigm in the investment. After he planted seeds, sophisticated flowers blossomed, such as the CAPM and APT, in the financial theory. In the real investment world, the Markowitz investment theories grew wealth and distributed it to investors and pension funds beneficially. Even in Japan, where there is a short history of employing modern finance theory in investment, there have been large benefits from his distinguished achievement.

In this paper, we discussed the efficiency of the market model in the real investment in Japanese equity. According to our results, there have been stable and statistical significant correlations between risk premium of the individual equities and the market portfolio. Therefore, we can conclude that Index fund investment or Market portfolio investment are efficient in 1990s. Furthermore, we examine the efficiency of the B/M, EP, and a scale proxy with the market model; however, we got only reliable result on the scale proxy. Before our estimation, we expected the book-to-market ratio (B/M) shows a same efficient result which will be a stable coefficient with positive sign. However, our result is not consistent with our hypothesis. We suggest that one of the reasons for unstable signs on coefficient of the B/M ratio is a conflict between legacy value and current market price. The implicit losses on the cross-shareholdings take time to reflect on the book. Therefore, the B/M ratio changed inelastically. This inelasticity causes loosing power of the B/M ratio to add an extra return on value stocks. We need to add do study for supporting our presumption under the mark-to-market basis after 2000. However, our suggestion will clear the puzzle for loosing power of the B/M ratio in 1990s in Japanese market.

References

Bernstein P.L. 1993. *Capital Ideas*. New York: Wiley, pp. 247–248.

Bloch, M., J.B. Guerard, H.M. Markowitz, and P. Todd. 1993. "Comparison of Some Aspects of U.S. and Japanese Equity Markets." *Japan and the World Economy* 5, pp. 3–28.

Chan, L., Y. Hamao, and J. Lakonishok. 1991. "Fundamental and Stock Returns in Japan." *Journal of Finance* 46, pp. 1739–1764.

Dimson, E. 1988. *Stock Market Anomalies*. Cambridge University Press, London.

Elton, E.J. and M.J. Gruber. 2007. *Modern Portfolio Theory and Investment Analysis*. New York: Wiley.

Fama, E.F. and K.R. French. 1992. "The Cross Section of Expected Stock Returns." *Journal of Finance* 47, pp. 427–465.

Fama, E.F. and K.R. French. 1996. "Multifactor Explanations of Asset Pricing Anomalies." *Journal of Finance* 51, pp. 55–84.

Fama, E.F. and K.R. French. 2002. "The Equity Premium." *Journal of Finance*, 57, No. 2, pp. 55–84.

Fama, E.F. and J.D. MacBeth. 1973. "Risk, Return and Equilibrium: Empirical Tests." *Journal of Political Economy* 81, pp. 607–636.

Guerard, J.B. 2006. "Quantitative Stock Selection in Japan and the United States: Some Past and Current Issues, *Journal of Investing* 15, No. 1, pp. 43–49.

Guerard, J.B., M. Takano, and Y. Yamane. 1993. "The Development of Mean–Variance Efficient Portfolios in Japan with Particular Emphasis on Sales and Earnings Forecasts." *Annals of Operations Research* 45, pp. 91–108.

Guerard, J.B., M. Suzuki, and B. Umesh. 1995. "Earnings Forecasting in Efficient Japanese Portfolios." in A. Chen, Ed., *Advances in Pacific Basin Business, Economics and Finance*, JAI Press Inc.

Konno, H. and H. Yamazaki. 1991. "Mean-Absolute Deviation Portfolio Optimization Model and its Application to Tokyo Stock Market". *Management Science* 37, pp. 519–531.

Kubota, K. and H. Takehara. 2007. "Re-examination for the Efficiency on Fama–French Factor Model." *Modern Finance* 22.

Lakonishok, J., A. Shleifer, and R.W. Vishny. 1994. "Contrarian Investment, Extrapolation and Risk." *Journal of Finance* 49, pp. 1541–1578.

Markowitz, H.M. 1952. "Portfolio Selection." *Journal of Finance* 7, pp. 77–91.

Markowitz, H.M. 1956. "The Optimization of a Quadratic Function Subject to Liner Constraints." *Naval Research Logistic Quarterly* 3, pp. 111–133.

Markowitz, H.M. 1976. "Investment for the Long Run: New Evidence for an Old Rule." *Journal of Finance* 36, pp. 1273–1285.

Markowitz, H.M. 1987. *Mean-Variance Analysis in Portfolio Choice and Capital Market*. Oxford: Basil Blackwell.

Pension Fund Association Japan. 1999. *Pension Fund Management under Deregulation on Investment*. Toyo Keizai.

Rubinstein, M. 2003. *A History of the Theory of Investments*. New York: Wiley.

Sharpe W.F. 1964. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *Journal of Finance* 19, pp. 425–442.

Watabe, H and T. Kobayashi. 2001."Estimated Earnings, Earnings Surprise and Value Effect," *Modern Finance* 9.

Wooten J.A. 2005. *The Employee Retirement Income Security Act of 1974*, University of California Press.

Chapter 27
Private Valuation of Contingent Claims in a Discrete Time/State Model

Alan J. King, Olga Streltchenko, and Yelena Yesha

27.1 Introduction

The no arbitrage paradigm provides the basis for modern financial theory and leads to the fundamental theorems of derivative pricing. In this theory, market prices for an underlying security and the related options tend to an equilibrium that can be characterized by a probability distribution on the future evolution of the underlying security price process. The first fundamental theorem states that a frictionless financial market is viable if and only if its future price evolution is compatible with an equivalent martingale measure called a *risk-neutral* measure. [Harrison and Pliska (1981) produced the first version of this theory in its modern form; see also the surveys of Duffie (1996) and Hull (2006).] The risk-neutral measure is used as a valuation operator: if a contract can be replicated in the underlying, then its price is an expected value of the payoffs stipulated by the contract with respect to a risk-neutral measure. In a complete market, existence of a unique pricing operator is guaranteed. For an incomplete market, uniqueness does not hold and many valuation operators are supported by the market.

The importance of risk-neutral measures for pricing generates interest in the inverse problem: given the observed market prices, retrieve a pricing measure compatible with them. This problem is known as *calibration* of a risk-neutral measure.

Approaches to calibration can be divided into parametric and nonparametric. The parametric approach assumes a particular distribution type and retrieves the parameters of the distribution. A contract is then analytically priced based on the chosen distribution for the risk-neutral measure; the market price of the observable contracts are substituted into the pricing formula, and the inverse problem is solved.

A.J. King (✉)
IBM Research, Yorktown Heights, New York, USA
e-mail: kingaj@us.ibm.com

O. Streltchenko
Bank of Montreal, Toronto, Ontario, USA

Y. Yesha
University of Maryland, Baltimore County, Maryland, USA August 29, 2009

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_27,
© Springer Science+Business Media, LLC 2010

The nonparametric approach allows for greater flexibility in the structure of the target distribution. The main idea is to use the most general no-arbitrage relations between a risk-neutral probability distribution and contract prices. For example, the price of a European contingent claim is an expected value of the payoffs at the time of maturity: $V_0 = E(F_T)$. According to its definition, the target probability measure is required to turn the price process of the underlying into a martingale. The market price of the contract is then substituted, and the inverse problem is solved.

In an incomplete market there is not enough information to completely define the system (i.e., to identify a unique risk neutral measure). Nonparametric calibration supplies a class of risk-neutral measures consistent with a given market. To choose among this class other methods are employed, such as kernel methods, curve fitting, and optimization methods. This chapter provides an intuitive, mathematical framework for calibration via optimization, and so we briefly review previous work in this area.

A comprehensive review of the calibration literature is provided in Jackwerth (1999). The topic of calibration in a discrete market was first studied by Rubinstein (1994). Market price movements were modeled with a binomial tree and the resulting risk-neutral probability measure obeyed an optimality condition for the mean square pricing error. Subsequent papers by Jackwerth and Rubinstein (1996) developed this approach to accommodate several expiration dates, different utility functions (optimization criteria), etc. The calibration of continuous probability measures via optimization is treated, among others by Avellaneda et al. (1997, 2001). Here, the distance function between the measures is chosen from an information theoretic point of view. Avellaneda et al. (1997) establishes the relative entropy of the target measure with respect to the prior as a calibration utility, and Samperi (1997) and Avellaneda et al. (2001) suggest the use of convex duality to connect calibration and portfolio optimization.

In this chapter, we explore further this convex duality argument to show that every portfolio optimization problem in a market that allows liquid trading is equivalent to a pricing measure retrieval problem and vice versa. According to our model, the resulting pricing measures reflect private valuation of the contingent claims by the investors. This result is a consequence of considering the portfolio optimization problem from the point of view of a hedger endowed with specific goals and liabilities. Such an investor operates in a "submarket" of available financial securities, which he sees as arbitrage-free. This perspective shifts our vision of no-arbitrage from a "global" to a "local" market, that is, to the one seen by an investor through the holdings of his/her portfolio.

The resulting mathematical framework allows one to explore the effect of the individual liability structure on the valuation operator. The goal of this chapter is to indicate the ways that optimization models of the management of claims can contribute to an understanding of how valuation operators should be formed to reflect an individual investor's view of the market. The advantage of the framework is that it stresses the individual view of the market by an investor and explicitely links preexisting liability structures of individual market participants to their valuation operators, as well as connecting investors' risk preferences (utility function) to the calibration utility.

This chapter treats the finite state theory, as surveyed recently by Naik (1995), Duffie (1996, Chap. 2), and Pliska (1997). It builds on the work presented in King (2001) and King et al. (2005). The primary contribution of our paper is the extension of the framework of King et al. (2005), Samperi (1997), and Avellaneda et al. (2001) to include asset-liability management problems and to stress the importance of liability structure as an incentive for trading.

The chapter proceeds as follows. Section 27.2 establishes terminology and notation for the discrete market and introduces some basic assumptions to be used throughout the paper. The main purpose of Sect. 27.3 is to introduce the portfolio optimization problem, so that we later recognize it as the dual to the calibration problem. Section 27.4 presents the APT point of view of the calibration problem, derives the problem dual to it and attaches a microeconomic model to the dual. Section 27.5 works in the other direction. We model an investor with a liability structure and develop the calibration problem corresponding to such an investor. Section 27.6 discusses some basic investor types with respect to their risk preferences and the calibration utilities corresponding to each of the types, and Sect. 27.7 discusses the practical implementation of a calibrated investment problem. Section 27.8 concludes with a discussion of calibration in view of the duality framework, significance of the calibration utility choice, and current and future research in this direction.

27.2 Discrete Market

To establish our terminology and notation we briefly recapitulate the discrete market formalism as presented in King (2001). The probability space is $(\Omega, \mathcal{F}, P)$ with Ω being finite and $P(\omega) > 0$ for every $\omega \in \Omega$. A community of investors agree on the possible states of the world but can have different preferences with respect to those states.

The securities are traded at fixed times $t = 0, 1, ..., T$ with T being the terminal date for market activities. There are $J + 1$ traded securities indexed by $j = 0, 1, ..., J$, and the vector price process $\{S_t\}$ generates a filtration $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq ...\mathcal{F}_t \subseteq \mathcal{F}_{t+1} \subseteq ... \subseteq \mathcal{F}_T = \mathcal{F}$. The true state of the world is revealed to the investors at time t by the atoms of $\mathcal{F}_t$. Every atom, or state $n \in \mathcal{F}_t$ except the initial one has a unique parent denoted $a(n)$, and every nonterminal state m has a set of child states $C(m)$. The set of atoms of $\mathcal{F}$ may, therefore, be viewed as nodes in a "scenario tree."

The probability distribution P is modeled by attaching weights $p_n > 0$ to each leaf node $n \in \mathcal{F}_T$ so that $\sum_{n \in \mathcal{F}_T} p_n = 1$. For each nonterminal node one has, recursively,

$$p_n = \sum_{m \in C(n)} p_m \quad \forall n \in \mathcal{F}_t, \quad t = T - 1, \ldots, 0$$

and so each node receives a probability mass equal to the combined mass of the paths passing through it. The ratios p_m / p_n, $m \in C(n)$, are the conditional probabilities that the child node m occurs given that the parent node $n = a(m)$ has occurred.

Following Harrison and Pliska (1981), let security 0 be the *numeraire* – a strictly positive process reflecting a riskless investment, like money-market, – and introduce the *discount process* $(1/S_t^0)$. Define the discounted price process $Z_t = S_t/S_t^0$. Note that $Z_n^0 = 1$ for every $n \in \mathcal{F}_t$ and every t. In what follows, all cash flows and prices are denominated in units of this riskless asset.

Investors have no influence on the prices of any security and may undertake trades at every time-step based on information accumulated up to time t. The amount of security j held by the investor in state $n \in \mathcal{F}_t$ is denoted θ_n^j. Let $\theta = \{\theta_t\}$ be an investment strategy or a series of portfolio decisions; then the (discounted) value of an investor's portfolio at any state n is

$$Z_n \cdot \theta_n := \sum_{j=0}^{J} Z_n^j \theta_n^j$$

Note that θ_t is predictable, that is, $\theta_t \in \mathcal{F}_{t-1}$, which means that an investor makes a decision about his/her portfolio holdings at time t before he observes the prices.

Our goal is to consider a community of investors that use contracts to protect themselves from market price fluctuations. Thus we would like to extend the strategy θ to include positions in the contracts available in the market. In the model described above there are $J+1$ basic securities whose evolution is governed by a given scenario tree. We extend the universe of tradable securities to include derivative securities whose payoffs depend on the values of the underlying $J+1$ basic securities.

In this chapter, we assume that there are only European-type securities in the market, meaning that the payoff dates are fixed for these contacts. The number of payoff dates is arbitrary but fixed, so we can assume that there is a payoff on each security, possibly equal to 0, at any $t = 1, \ldots, T$.

27.3 Optimization Model for an Investor with an Unconstrained Portfolio

In this discrete market of basic securities and derivative contracts written on them, investors consider decisions on whether or not to enter a specific financial contract at a given price at each time stage $t = 0, 1, \ldots, T$. Suppose such an investor considers buying or selling a contract at time $t = 0$. The investor's problem is to maximize his/her utility function (of the terminal wealth) while taking the following into consideration. If a contract is sold, the investor needs to obtain enough money from the sale to invest it in the market of basic securities and hedge the payoffs on the contract without risk of falling short. If a contract is bought, then the cash flows it generates should serve to hedge the investor's exposure to the fluctuations in the price of the $J+1$ basic assets. The trades therefore must be self-financing: the portfolio value after trading at time t equals the portfolio value before trading minus the required payoffs.

As is common in microeconomic modeling, we assume the utility functions of the investors to be strictly increasing and to have their domain equal to or contained in the nonnegative reals:

$$f(w) = \begin{cases} \tilde{f}(w) & \text{for } w \geq 0 \\ -\infty & \text{for } w < 0, \end{cases} \tag{27.1}$$

where w is the investor's terminal wealth. This essentially means that 0 is an absorbing state of the decision system. We further assume that there are no borrowing, short-selling or budget constraints imposed on the investor. The reason for this assumption will become clear after our discussion of the calibration problem in Sect. 27.4. We will see that the problem dual to the calibration problem (27.5) gives rise to a portfolio optimization problem with these properties.

The initial wealth is equal to the initial endowment V_0 plus the amount received from the purchase or sale of the contract zF_0, where z represents the quantity. Note that if the contract is sold then $zF_0 \geq 0$ and if a contract is bought then $zF_0 \leq 0$. This amount, $V_0 + zF_0$, is invested in the initial portfolio of basic securities $Z_0\theta_0$.

In the notation introduced above the self-financing condition is

$$Z_n\theta_n = Z_n\theta_{a(n)} - zF_n, \; n \in \mathcal{F}_t, \; t > 0,$$

where zF_n is a discounted payoff on the contract at a state $n \in \mathcal{F}_t$. We assume that the contract price F_0 is quoted at the market. Given the price F_0, the above stated investor's decision problem can be formalized as follows:

$$\begin{aligned} \max_{(\theta,z)} \quad & E^P(f(Z_T\theta_T)) \\ & Z_0\theta_0 - V_0 - zF_0 = 0 : y_0 \\ & Z_n[\theta_n - \theta_{a(n)}] + zF_n = 0, \; n \in \mathcal{F}_t, \; 1 \leq t \leq T : y_n. \end{aligned} \tag{27.2}$$

If several contracts are considered to be traded, the term zF_0 should be replaced by a summation $\sum_{i\in\mathcal{K}} z_i F_{i0}$, and zF_n should be replaced by $\sum_{i\in\mathcal{K}} z_i F_{in}$.

27.4 Calibration of a Discrete Market

In this section, we will look at the problem of calibration as retrieval of a pricing measure compatible with all the contracts in the market, develop its dual problem, and compare it to the portfolio optimization problem (27.2). Such analysis will be instrumental in understanding the underlying microeconomic models and the choice of meaningful calibration utilities.

Let $\mathcal{K}$ be a collection of derivative securities indexed by $i = 1, \ldots, M$, and let C_i denote the price of the ith security in $\mathcal{K}$. Let $\mathcal{Q}$ be a collection of risk-neutral measures compatible with all the instruments in the market, and Q be a member of this set. The price processes for the underlying securities is a martingale with respect to Q by definition of a risk-neutral measure. According to APT, the same property extends to European-type derivatives (e.g., Karatzas and Shreve 1998). Thus the prices of the European-type derivatives are expected values of their future payoffs with respect to the target risk-neutral measure. The observable information consists of a set of security prices at time 0. A risk-neutral measure consistent with the contract prices must obey

$$\begin{aligned} \textstyle\sum_{m \in C(n)} q_m Z_m &= q_n Z_n,\ n \in \mathcal{F}_t,\ 0 \le t \le T-1, \\ \textstyle\sum_{t=1}^{T} \sum_{n \in \mathcal{F}_t} F_{in} q_n &= C_i, \quad 1 \le i \le M, \\ q_n &\geqslant 0, \\ q_0 &= 1. \end{aligned} \tag{27.3}$$

Notice that $Z_n^0 = 1$ implies that the q_n are the weights of a martingale probability measure Q. It is convenient also to use the notation q to represent the measure; one can think of q as being the representation of the probability Q over the scenario tree.

Equations in (27.3) are called *calibration equations*. When this system is underdefined, the problem is viewed as one of incomplete information. An optimization criterion helps choose among the members of $\mathcal{Q}$. There are only finitely many states in the discrete case, so assigning finite probabilities to each state results in a finite measure. The optimization criterion is often interpreted as a "distance" $g(q, p)$ from some prior measure p, where $g(\cdot, p)$ is convex. Therefore, we will incorporate p as a parameter into the distance measure, and for convenience we also incorporate the nonnegativity and scaling conditions into the optimization criterion:

$$g(q, p) = \begin{cases} g(q, p) & \text{for } q \ge 0 \text{ and } q_0 = 1 \\ \infty & \text{otherwise.} \end{cases} \tag{27.4}$$

The calibration problem can now be stated as an optimization problem as follows

$$\begin{aligned} &\min_{(q)} \quad \sum_{n \in \mathcal{F}_T} g(q_n, p_n) \\ &\qquad q_n Z_n - \textstyle\sum_{m \in C(n)} q_m Z_m = 0,\ n \in \mathcal{F}_t,\ t \le T-1 : \omega_n \\ &\qquad C_i - \textstyle\sum_{t=1}^{T} \sum_{n \in \mathcal{F}_t} F_{in} q_n = 0,\ i \in \mathcal{K} : z_i \end{aligned} \tag{27.5}$$

(In a discrete probability setting, the optimization criterion needs only reference to the terminal states of the scenario, because these weights determine the probability at all intermediate nodes.) The following theorem shows that the calibration problem is dual to a certain portfolio investment problem in which the investor is permitted to take initial positions in the derivatives, and in which the "utility" of the investor is related to the convex conjugate $g^*(w) = g^*(w, p)$, where

$$g^*(w, p) = \sup_q \{wq + g(q, p)\}. \tag{27.6}$$

A similar result was achieved in Samperi (1997).

Theorem 27.1. *The calibration problem (27.5) is equivalent to the portfolio optimization problem*

$$\begin{aligned} &\max_{(\theta,z)} \sum_{n\in\mathcal{F}_T} h(Z_n\theta_n, p_n) \\ &\quad Z_0\theta_0 - \sum\nolimits_{i\in\mathcal{K}} z_i C_i = 0 \\ &\quad Z_n[\theta_n - \theta_{a(n)}] + \sum\nolimits_{i\in\mathcal{K}} z_i F_{in} = 0,\ n \in \mathcal{F}_t,\ 1 \le t \le T \end{aligned} \tag{27.7}$$

where for each p the function $h(\cdot, p)$ is the reflection through the origin of the convex conjugate of the calibration distance

$$h(w, p) = -g^*(-w, p).$$

Proof. See Appendix 27.9. □

Note that for strong duality between (27.5) and (27.7) to hold the affine inequalities [embedded in (27.4)] do not need to hold with strict inequality since the optimization problem is convex (see Boyd and Vandenberghe 2008, pp. 226–227 for discussion).

The main differences between the portfolio optimization problem (27.2) and the calibration dual (27.7) are the form of the utility function and the domain of the investment universe. First, the calibration objective function $h(Z_n\theta_n, p_n)$ is not necessarily expressible in a form $\hat{f}(Z_n\theta_n)p_n$, so the expectation $\sum_{n\in\mathcal{F}_T} h(Z_n\theta_n, p_n)$ is not necessarily an expected utility in the form $E^P(\hat{f}(Z_N\theta_n))$. We will give an example of such an outcome in Sect. 27.6. Further comparing (27.7) with (27.2), one concludes that the investor represented by (27.7) considers taking position in all the contracts available in the market and exercising buy-and-hold strategy toward the derivative securities.

27.5 Investor with Liabilities and Endowments

Suppose an investor wishes to hedge a liability stream (which may not necessarily be a liquid security). As in King (2001), existing liability structures serve as an incentive for the investor to consider trading in contracts offered by the market. We will now explore the influence of this liability structure on the calibration of the risk-neutral measure. In effect, we will retrieve a pricing operator consistent with the investor's portfolio and utility function.

An existing liability structure or endowment is represented in our model the following way: in each state n there is a payoff (positive or negative) denoted L_n – similar to payoffs of a contigent claim. Although the liability may not itself be a tradeable security, the investor's ability to make corresponding payments is tied to his/her market performance. As before, the initial endowment of the investor is denoted V_0. The investor's decision problem becomes to invest in the market of liquid securities to hedge all the paoyoffs on his/her liabilities, including those generated by entering market contracts, without falling short. This model can be applied to calibrate the pricing measure for an investor taking a position in a thinly traded contract or custom derivative.

We incorporate the liability structure and the available liquid claims into the portfolio optimization problem:

$$\begin{aligned}
&\max_{(\theta,z)} E^P(f(Z_T\theta_T)) \\
&\quad Z_0\theta_0 - \sum\nolimits_{i\in\mathcal{K}} z_i F_{i0} = V_0 : y_0, \qquad (27.8)\\
&\quad Z_n[\theta_n - \theta_{a(n)}] + \sum\nolimits_{i\in\mathcal{K}} z_i F_{in} = -L_n,\ n\in\mathcal{F}_t,\ t\geq 1 : y_n.
\end{aligned}$$

Note that the investor is required to posses a nonnegative terminal wealth, therefore the utility $f(w)$ is of the form (27.1). Next, let us introduce the concave conjugate function $f^*(\cdot)$ of the investor's utility function $f(\cdot)$ by

$$f^*(y) = \inf_w\{yw - f(w)\}. \qquad (27.9)$$

The following theorem is the main result of this paper.

Theorem 27.2. *The investor's decision problem (27.8) is equivalent to the liability-modified calibration problem*

$$\begin{aligned}
&\min_{q,\,y_0} y_0\left(V_0 - \sum_{t=1}^{T}\sum_{n\in\mathcal{F}_t} L_n q_n\right) - \sum_{n\in\mathcal{F}_T} f^*(y_0 q_n/p_n)p_n \\
&\quad Z_n q_n - \sum\nolimits_{m\in C(n)} q_m Z_m = 0,\ n\in\mathcal{F}_t,\ 0\leq t\leq T \\
&\quad F_{i0} - \sum_{t=1}^{T}\sum\nolimits_{n\in\mathcal{F}_t} F_{in}q_n = 0,\ i\in\mathcal{K} \qquad (27.10)\\
&\quad q_0 = 1,\\
&\quad y_0 > 0.
\end{aligned}$$

Proof. See Appendix 27.10. □

Similarly to the calibration problem case (Sect. 27.4) strong duality holds here too.

In this result, we see that the investor's initial capital and liability structure do not contribute to the constraints of the dual problem, but establish the linear part of the utility function to be minimized. Observe that for problem (27.11) to be bounded, and thus for (27.8) to be feasible,

$$V_0 \geqslant \sum_{t=1}^{T} \sum_{n \in \mathcal{F}_t} L_n q_n$$

needs to hold. Therefore, the minimal initial capital required to fund the liability in the market is

$$V_0 := \sum_{t=1}^{T} \sum_{n \in \mathcal{F}_t} L_n q_n.$$

This sets V_0 to be the risk-neutral discounted value of the liability with respect to the calibrated measure Q. In applications where V_0 plays the role of actual investment capital, a utility function that enforces nonnegative wealth could lead to infeasibility. In such applications, one will have to assess whether negative wealth outcomes are reasonable for the given situation.

The nonlinear part of the calibration utility is a conjugate function of the original investor's utility with P becoming the prior for the calibrated measure Q.

In the next section, we explore some varieties of risk preferences and their corresponding calibration kernels.

27.6 Calibration and Utilities for Various Types of Risk Preferences

In this section, we will explore the connection between some widely used utility functions expressing investor's risk attitude and the resulting calibration utilities. One of the most popular portfolio optimization utilities yields a calibration utility that seems to have been neglected by the calibration research community. We will also see an example of a distance function between the target risk-neutral measure and the prior that does not yield a meaningful utility function for the investor's portfolio optimization problem. This suggests that the choice of the calibration utility should be informed by the notion of the duality between the two problems: portfolio optimization and calibration. The utility functions discussed here represent risk-averse investors with different absolute and relative risk aversion Elton and Gruber (1991).

Consider an investor with constant absolute risk aversion expressed through the following utility:

$$f(w) = -\mathrm{e}^{w}.$$

A standard calculation (see Rockafellar 1972, p. 105) shows that

$$\begin{aligned} f^*(x) &= \inf_w [xw + \mathrm{e}^{-w}] \\ &= x - x \ln x, \end{aligned}$$

which, after observing that $\sum_{n \in \mathcal{F}_T} q_n = 1$, contributes the following terms to the calibration problem (27.11)

$$\sum_{n \in \mathcal{F}_T} q_n \ln \frac{q_n}{p_n} + y_0 \ln y_0 - y_0.$$

The first term is known as the *relative entropy* of q with respect to p. Relative entropy has been widely employed as a calibration utility (e.g., Avellaneda et al. 1997). The correspondence between the exponential portfolio optimization utility and the relative entropy was recognized in Avellaneda et al. (2001). On the basis of the above discussion, we may conclude that calibration using relative entropy of the target risk-neutral measure with respect to some prior probability measure is equivalent to portfolio optimization for an investor with constant absolute risk aversion. Notice, however, that this pair of utility/calibration kernels does not enforce nonnegative wealth constraints in the utility formulation of (27.8).

Next consider an investor with decreasing absolute risk aversion and constant relative risk aversion. Such an investor has a utility of the form

$$f(w) = \ln w.$$

Then

$$f^*(x) = 1 + \ln x,$$

and the contribution to the calibration problem becomes

$$-\sum_{n \in \mathcal{F}_T} p_n \ln \frac{p_n}{q_n} - 1 - \ln y_0,$$

where the first term is the relative entropy of the prior p with respect to q. To our knowledge, this version of relative entropy has not been investigated in the literature. It could be interesting for applications, because in contrast to the exponential utility, the log-utility does enforce nonnegative wealth.

Another simple utility of interest is a quadratic utility

$$f(w) = w - bw^2,$$

where b is positive. It represents an investor with increasing absolute and relative risk aversion, such as a mean–variance investor Markowitz (1952). For this utility, the conjugate works out to be

$$f^*(x) = -\frac{(1-x)^2}{4b}.$$

The contribution to the calibration problem is

$$\frac{1}{4b} \sum_{n \in \mathcal{F}_T} \frac{(y_0 q_n - p_n)^2}{p_n}.$$

This calibration kernel is encountered in the calibration literature (with $b = 1/4$ and $y_0 = 1$) under the name of *goodness of fit* Jackwerth and Rubinstein (1996):

$$\sum_{n \in \mathcal{F}_T} \frac{(q_n - p_n)^2}{p_n}.$$

Thus, we see that the corresponding calibration kernel for the mean–variance investment framework is the goodness of fit.

In many situations it may be the case that the investor would like to enforce nonnegative terminal wealth. This terminal condition will change the calibration problem, of course. The following theorem shows how the calibration problem is affected.

Theorem 27.3. *Consider the investor's decision problem* (27.8) *with the additional constraint that terminal wealth be nonnegative:*

$$Z_n \theta_n \geq 0, \quad n \in \mathcal{F}_T. \tag{27.11}$$

Then the dual calibration problem is

$$\begin{aligned}
\min_{q,\, y_0} y_0 \left(V_0 - \sum_{t=1}^{T} \sum_{n \in \mathcal{F}_t} L_n q_n \right) - \sum\nolimits_{n \in \mathcal{F}_T} f^* \left(y_0 \frac{q_n + t_n}{p_n} \right) p_n \\
Z_n q_n - \sum\nolimits_{m \in C(n)} q_m Z_m = 0,\ n \in \mathcal{F}_t,\ 0 \leq t \leq T \\
F_{i0} - \sum_{t=1}^{T} \sum_{n \in \mathcal{F}_t} F_{in} q_n = 0,\ i \in \mathcal{K} \\
t_n \geq 0,\ n \in \mathcal{F}_T \\
q_0 = 1, \\
y_0 > 0.
\end{aligned} \tag{27.12}$$

Proof. See Appendix 27.11. □

The interpretation of this dual problem is standard. The imposition of additional constraints on the primal creates new degrees of freedom in the dual. If a given scenario n had a negative terminal wealth at optimality, then the imposition of the nonnegativity condition will likely force terminal wealth along that scenario n to zero and the corresponding dual quantity t_n will also likely be strictly positive. This has an initial effect of lowering the objective value through the influence of t_n in the term $f^*(y_0\frac{q_n+t_n}{p_n})$. The optimization now has the opportunity to shift q_n higher along scenario n and so has an opportunity to move the risk-neutral probability mass around to achieve even lower values of the optimum. Of course, in this process the value of the integral $\sum_{t=1}^{T}\sum_{n\in\mathcal{F}_t} L_n q_n$ will likely also change but this will depend on the precise tradeoffs modeled in the calibration kernel.

Finally, we would like to demonstrate how understanding of portfolio optimization and calibration duality becomes instrumental when choosing a calibration utility. Suppose the distance between the target risk-neutral measure and the prior distribution is chosen to be the squared L_2-distance on the atoms of $\mathcal{F}_T$ and the calibration utility becomes

$$\sum_{n\in\mathcal{F}_T}(y_n - p_n)^2.$$

This calibration utility was, for example, used in Rubinstein (1994).

Taking the conjugate, one obtains $f^*(w_n) = w_n^2/4 + w_n p_n$, where $w_n = -Z_n\theta_n$ (see Sect. 27.4), and the portfolio optimization utility becomes

$$\sum_{n\in\mathcal{F}_T}[Z_n\theta_n p_n - (Z_n\theta_n)^2] = E^P(Z_n\theta_n) - \sum_{n\in\mathcal{F}_T}(Z_n\theta_n)^2.$$

While the first term is in the form of an expectation with respect to the prior, the second term can be interpreted as a scaled expectation with respect to the uniform distributions on the atoms of $\mathcal{F}_T$. The possibility of such an outcome was mentioned in Sect. 27.4. Since this utility cannot be expressed in the form of an expectation of the terminal wealth with respect to P such a utility would not normally be chosen to represent investor's preferences. Therefore, the use of the squared L_2-distance as a calibration kernel cannot be justified from the utility perspective we are using here.

27.7 The Calibrated Investment Problem

The discussion up to this point has primarily focussed on the relationship between calibration and utility. In this section, we discuss some practical matters related to the implementation of the basic framework of Theorem 27.2.

It is a fact that an attempt to solve the system of calibration equations (27.3) using posted bid-ask spreads in options markets will normally fail to find a feasible solution. In theory this means that prices in the market almost always allow arbitrage trading strategies! Of course, this statement cannot be taken too seriously. The actively traded options will have current prices that are practically arbitrage-free, and the inactive options probably have out-of-date prices that may allow for some arbitrage. Market- makers protect themselves from arbitrage by offering very low volumes in their quotes for these inactive options. The instant someone starts to trade in these options, the market-makers will notice this spike in volume and adjust their quotes to the current state of the market. What this means in practice is that an implementation of the framework of Theorem 27.2 must take into account this basic point concerning volumes.

The simplest mechanism to ensure the feasibility of the calibration equations is to relax the equality constraints for the options prices and incorporate a penalty term into the calibration objective. If we adopt the volumes V_i as the penalty multipliers, then the pair of liquidity-adjusted utility/calibration models are as follows. (We ignore the issue of bid-ask spreads and set the quote to be the midpoint.) The calibration problem then adds a penalty term in the pricing error for the option price quotes:

$$
\begin{aligned}
\min_{q,\, y_0} y_0 \left(V_0 - \sum_{t=1}^{T} \sum_{n \in \mathcal{F}_t} L_n q_n + \sum_i V_i (u_i^+ + u_i^-) \right) - \sum\nolimits_{n \in \mathcal{F}_T} f^*(y_0 q_n / p_n) p_n & \\
Z_n q_n - \sum\nolimits_{m \in C(n)} q_m Z_m &= 0,\ n \in \mathcal{F}_t,\ 0 \le t \le T \\
F_{i0} - \sum_{t=1}^{T} \sum_{n \in \mathcal{F}_t} F_{in} q_n &= u_i^+ - u_i^-,\ i \in \mathcal{K} \\
u_i^+, u_i^- &\geqslant 0 \\
q_0 &= 1, \\
y_0 &> 0.
\end{aligned}
\tag{27.13}
$$

The corresponding utility maximization problem is:

$$
\begin{aligned}
\max_{(\theta, z)} \ & E^P(f(Z_T \theta_T)) \\
& Z_0 \theta_0 - \sum\nolimits_{i \in \mathcal{K}} z_i F_{i0} = V_0 : y_0 \\
& Z_n [\theta_n - \theta_{a(n)}] + \sum\nolimits_{i \in \mathcal{K}} z_i F_{in} = -L_n,\ n \in \mathcal{F}_t,\ t \ge 1 : y_n \\
& z_i \in [-V_i, V_i].
\end{aligned}
\tag{27.14}
$$

We see in the utility maximization problem that the volume is indeed the natural penalty term to use, since the imposition of a penalty in the calibration introduces bounds on the static positions in the option instruments in the investment problem. The system (27.14) is called the *calibrated investment problem*.

A final point concerns the application of the calibrated investment problem as a practical tool for investment management. The dual solution q is a representation of the risk-neutral measure appropriate for the given liability structure and the market for underlying securities and options. It provides the manager with a *private* risk-neutral valuation operator. The dual solution will be a reasonable approximation of the sensitivity of the impact of *small changes* to the liability structure. Using the terminology of convex analysis, the risk-neutral valuation operator is a local subgradient of the expected utility considered as a functional of the liability. A formal expression of this relationship is as follows:

$$\nabla_L \left(E^P f(Z_T \theta_T; L) \right) (\delta L) = E^Q \delta L. \tag{27.15}$$

Such formal relationships can be derived even when the underlying calibrated investment problem is formulated in continuous time; see Rockafellar (1974) for a survey of the mathematical duality required for this kind of analysis. As a practical matter, most applications will adopt a linear or a mean–variance framework for the calibrated investment problem. These problems can be solved quite readily by commercially available and even open source Foundation (2008) linear and quadratic programming solvers, which as a matter of course will produce the dual variables q. These dual variables can then be used to produce a *private valuation* of small changes to the liability structure, in the sense that

$$\sum_{t=1}^{T} \sum_{n \in \mathcal{F}_t} q_n \delta L_n \tag{27.16}$$

will be a very good approximation to the impact on the expected utility objective to a small change δL_n in the underlying liability structure, such as might be caused by an over-the-counter transaction or other type of change to the liability stream.

27.8 Conclusions

The duality framework discussed in this chapter illustrates some basic relationships between calibration and utility. Since an investor's portfolio properties, risk preferences, liability structure, and the segment of the market considered for trading activity determine his/her private valuation operator, then to price a new security against an established market one should explicitly include the new liability structure that arises as a result of the investor taking positions in this security.

The liability generates a linear adjustment term in the resulting calibration problem, while the nonlinear part of the calibration utility is given by a conjugate function of the original investor's utility. The probability measure believed by the investor to be the physical probability measure underlying the market evolution becomes what is commonly known as the prior for the resulting calibration problem and affects the calibration utility through its nonlinear term.

Understanding this duality helps to choose a meaningful calibration utility. The portfolio optimization problem is amenable to modeling and yields a utility function through some basic microeconomic considerations, while calibration is a more abstract problem that does not directly rely on microeconomic arguments for its structure. Thus, it makes sense to check whether the calibration utility to be used yields a meaningful microeconomic model in the implied portfolio optimization problem.

A calibrated measure obtained through the traditional APT approach cannot be used to price a new liability stream, for example, a new contract being introduced into the market, etc., since the liability structure changes the calibration utility. However, once the liability structure is explicitly included in the investor's decision problem, the solution to such a system will always result in a correct pricing operator along with a hedging strategy.

The approach proposed in Sect. 27.7 is a practical framework for pricing custom liability streams against an established market of liquid securities. Research topics that are suggested by this chapter include: incorporation of optional exercise features in the modeling of the calibration and liability streams, and investigation of calibration of multidimensional liability structures. For an excellent review of modeling issues in calibration, see Cont (2006)

27.9 Proof of Theorem 27.1

The theorem depends on showing that the two problems (27.5) and (27.7) are dual problems. The development of the duality argument follows the general patterns in Rockafellar (1974). To develop the problem dual to (27.5), the constraints are multiplied by the dual variables in the right margin of (27.5) and the Lagrangian is formed:

$$
\begin{aligned}
L(q, \omega, z) = & \sum_{n \in \mathcal{F}_T} g(q_n, p_n) \\
& - \sum_{t=0}^{T-1} \sum_{n \in \mathcal{F}_t} \omega_n \left(q_n Z_n - \sum_{m \in C(n)} q_m Z_m \right) \\
& - \sum_{i \in \mathcal{K}} z_i \left(C_i q_0 - \sum_{t=1}^{T} \sum_{n \in \mathcal{F}_t} F_{in} q_n \right).
\end{aligned}
$$

Collecting terms involving q_n:

$$
\begin{aligned}
L(q, \omega, z) = \sum_{n \in \mathcal{F}_T} &\left[g(q_n, p_n) + q_n \left(Z_n \omega_{a(n)} + \sum_{i \in \mathcal{K}} z_i F_{in} \right) \right] \\
&- q_0 \left[Z_0 \omega_0 + \sum_{t=1}^{T} \sum_{i \in \mathcal{K}} z_i C_i \right] \\
&- \sum_{t=1}^{T-1} \sum_{n \in \mathcal{F}_t} q_n \left[Z_n \left(\omega_n - \omega_{a(n)} \right) - \sum_{i \in \mathcal{K}} z_i F_{in} \right].
\end{aligned}
$$

The dual problem is then achieved by taking $\inf_{(q)} L(q, \omega, z)$. The constraints of the dual are set by the requirement that the factors of q_n must evaluate to zero for a feasible dual solution:

$$
\begin{aligned}
\max_{(\omega, z)} \; & \sum_{n \in \mathcal{F}_T} \left(-g^* \left(-Z_n \omega_{a(n)} - \sum_{i \in \mathcal{K}} z_i F_{in}, p_n \right) \right) \\
& Z_0 \omega_0 + \sum_{i \in \mathcal{K}} z_i C_i = 0, \\
& Z_n [\omega_n - \omega_{a(n)}] - \sum_{i \in \mathcal{K}} z_i F_{in} = 0, \; n \in \mathcal{F}_t, \; 1 \le t \le T-1, \quad (27.17)
\end{aligned}
$$

where $g^*(w_n, q_n)$ is the convex conjugate of $g(q_n, p_n)$ in the first variable:

$$
g^*(w, p) = \sup_q [wq - g(q, p)].
$$

Note that from the second constraint of (27.17) it follows that

$$
-Z_n \omega_{a(n)} - \sum_{i \in \mathcal{K}} z_i F_{in} = -Z_n \omega_n.
$$

Perform a variable change $\theta = -\omega$, let $h(-w, p) = -g^*(w, p)$ and rewrite (27.17) to obtain (27.7).

27.10 Proof of Theorem 27.2

This proof can be derived from the results in (King, 2001, Sect. 8). Here, we give the explicit proof. As above we construct the problem dual to (27.8) following the general approach of Rockafellar (1974). The constraints are multiplied by the dual variables in the right margin of (27.8), and the Lagrangian is formed:

$$
\begin{aligned}
L(\theta, z, y) = &\sum_{n \in \mathcal{F}_T} f(Z_n \theta_n) p_n \\
&- y_0 \left(Z_0 \theta_0 - V_0 - \sum_{i \in \mathcal{K}} z_i F_{i0} \right) \\
&- \sum_{t=1}^{T} \sum_{n \in \mathcal{F}_t} y_n \left(Z_n \left[\theta_n - \theta_{a(n)} \right] + L_n + \sum_{i \in \mathcal{K}} z_i F_{in} \right) \\
= &\sum_{n \in \mathcal{F}_T} \left[f(Z_n \theta_n) p_n - y_n Z_n \theta_n \right] \\
&- y_0 \left(Z_0 \theta_0 - V_0 - \sum_{i \in \mathcal{K}} z_i F_{i0} \right) \\
&- \sum_{t=1}^{T-1} \sum_{n \in \mathcal{F}_t} y_n \left(Z_n \left[\theta_n - \theta_{a(n)} \right] + L_n + \sum_{i \in \mathcal{K}} z_i F_{in} \right) \\
&- \sum_{n \in \mathcal{F}_T} y_n \left(-Z_n \theta_{a(n)} + L_n + \sum_{i \in \mathcal{K}} z_i F_{in} \right).
\end{aligned}
$$

After the factors of θ_n and z are collected, and noting that $p_n > 0$:

$$
\begin{aligned}
L(\theta, z, y) = &\sum_{n \in \mathcal{F}_T} \left[f(Z_n \theta_n) - \frac{y_n}{p_n} Z_n \theta_n \right] p_n \\
&- \sum_{t=1}^{T} \sum_{n \in \mathcal{F}_t} y_n L_n + y_0 V_0 \\
&+ \sum_{i \in \mathcal{K}} z_i \left(y_0 F_{i0} - \sum_{t=1}^{T} \sum_{n \in \mathcal{F}_t} y_n F_{in} \right) \\
&- \sum_{t=1}^{T-1} \sum_{n \in \mathcal{F}_t} \theta_n \left(y_n Z_n - \sum_{m \in C(n)} y_m Z_m \right).
\end{aligned}
$$

The dual problem is then achieved by taking $\sup_{(\theta,z)} L(\theta, z, y)$. The constraints of the dual are set by the requirement that the factors of θ_n and z must evaluate to zero for a feasible dual solution:

$$
\begin{aligned}
\min_{y} \quad & y_0 V_0 - \sum_{t=1}^{T} \sum_{n \in \mathcal{F}_t} y_n L_n - \sum_{n \in \mathcal{F}_T} f^*(y_n / p_n) p_n \\
& Z_n y_n - \sum_{m \in C(n)} y_m Z_m = 0, \; n \in \mathcal{F}_t, \; 1 \le t \le T,
\end{aligned}
$$

$$y_0 F_{i0} - \sum_{t=1}^{T} \sum_{n \in \mathcal{F}_t} y_n F_{in} = 0, \; i \in \mathcal{K}, \tag{27.18}$$

where $f^*(\cdot)$ is the concave conjugate of $f(\cdot)$:

$$f^*(y) = \inf_w \big[yw - f(w) \big].$$

Since $f(\cdot)$ is assumed to be strictly increasing, the domain of its conjugate $f^*(\cdot)$ is bounded below by a positive constant, say $c > 0$. It follows that $y_n > cp_n$ for all $n \in \mathcal{F}_T$. From the martingale equation for the numeraire asset we have $y_0 = \sum_{n \in \mathcal{F}_T} y_n$, hence $y_0 > 0$. Therefore , let $q_n = \frac{y_n}{y_0}$, and rewrite (27.18) to obtain (27.11).

27.11 Proof of Theorem

27.3 The steps of the proof are identical to Appendix 27.10, except we must now incorporate nonnegativity constraints in the primal:

$$\begin{aligned}
\max_{(\theta,z)} \quad & E^P(f(Z_T \theta_T)) \\
& Z_0 \theta_0 - \sum_{i \in \mathcal{K}} z_i F_{i0} = V_0 : y_0 \\
& Z_n[\theta_n - \theta_{a(n)}] + \sum_{i \in \mathcal{K}} z_i F_{in} = -L_n, \; n \in \mathcal{F}_t, \; t \geq 1 : y_n \\
& Z_n \theta_n \geq 0, \; n \in \mathcal{F}_T : s_n
\end{aligned}$$

The new dual variables for the terminal wealth constraints are indicated on the right column. As in Appendix 27.10, we form the Lagrangian and collect primal terms. The dual variables s_n are required to be nonnegative, and the Lagrangian terms are required to have a negative sign. (The variables s_n are minimization variables. A negative value of $Z_n \theta_n$ will force the dual objective to plus infinity, generating a constraint for the primal that forces $Z_n \theta_n \geq 0$.)

$$\begin{aligned}
L(\theta, z, y, s) = & \sum_{n \in \mathcal{F}_T} f(Z_n \theta_n) p_n \\
& - \sum_{n \in \mathcal{F}_T} s_n Z_n \theta_n \\
& - y_0 \left(Z_0 \theta_0 - V_0 - \sum_{i \in \mathcal{K}} z_i F_{i0} \right) \\
& - \sum_{t=1}^{T} \sum_{n \in \mathcal{F}_t} y_n \left(Z_n \big[\theta_n - \theta_{a(n)} \big] + L_n + \sum_{i \in \mathcal{K}} z_i F_{in} \right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{n \in \mathcal{F}_T} \left[f(Z_n \theta_n) p_n - (y_n + s_n) Z_n \theta_n \right] \\
&\quad - y_0 \left(Z_0 \theta_0 - V_0 - \sum_{i \in \mathcal{K}} z_i F_{i0} \right) \\
&\quad - \sum_{t=1}^{T-1} \sum_{n \in \mathcal{F}_t} y_n \left(Z_n \left[\theta_n - \theta_{a(n)} \right] + L_n + \sum_{i \in \mathcal{K}} z_i F_{in} \right) \\
&\quad - \sum_{n \in \mathcal{F}_T} y_n \left(-Z_n \theta_{a(n)} + L_n + \sum_{i \in \mathcal{K}} z_i F_{in} \right).
\end{aligned}
$$

Following the rest of the steps in Appendix 27.10 and letting $t_n = \frac{s_n}{y_0}$ completes the proof of Theorem 27.3.

References

M. Avellaneda, R. Buff, C. Friedman, N. Grandchamp, L. Kruk, and J. Newman, *Weighted monte-carlo: A new technique for calibrating asset-pricing models*, International Journal of Theoretical and Applied Finance, 4 (2001), pp. 91–119.

M. Avellaneda, C. Friedman, R. Holmes, and D. Samperi, *Calibrating volatility surfaces via relative entropy minimization*, Applied Mathematical Finance, 4 (1997), pp. 37–64.

S. Boyd and L. Vandenberghe, *Convex Optimization*, 6 ed., Cambridge University Press, Cambridge, 2008.

R. Cont, *Model uncertainty and its impact on the pricing of derivative instruments*, Mathematical Finance, 16 (2006), pp. 519–547.

D. Duffie, *Dynamic Asset Pricing Theory*, Princeton University Press, Princeton, New Jersey, 1996.

E. J. Elton and M. J. Gruber, *Modern Portfolio Theory and Investment Analysis*, Wiley, New York, 1991.

C.-O. Foundation, *Computational infrastructure for operations research.* http://www.coin.org/, 2008.

J. M. Harrison and S. R. Pliska, *Martingales and stochastic integrals in the theory of continuous time trading*, Stochastic Processes and their Applications, 11 (1981), pp. 215–260.

J. Hull, *Options, Futures, and Other Derivatives*, 6 ed., Prentice Hall, Englewood, Cliffs, New Jersey, 2006.

J. C. Jackwerth, *Option implied risk-neutral distributions and implied binomial trees: A literature review*, The Journal of Finanace, 7 (1999), pp. 66–82.

J. C. Jackwerth and M. Rubinstein, *Recovering probability distributions from option prices*, The Journal of Finanace, 51 (1996), pp. 1611–1631.

I. Karatzas and S. Shreve, *Methods of Mathematical Finance*, Springer-Verlag, New York, New York, 1998.

A. King, *Duality and martingales: A stochastic programming perspective on contingent claims*, Mathematical Programming, (2001). Published online; available at http://link.springer.de/link/service/journals/10107/first/bibs/s101070100257.htm.

A. King, M. Koivu, and T. Pennanen, *Calibrated option bounds*, International Journal Of Theoretical and Applied Finance, 8 (2005), pp. 141–159.

H. MARKOWITZ, *Portfolio Selection: Efficient Diversification of Investments*, Wiley, New York, 1952.

V. NAIK, *Finite state securities market models and arbitrage*, in Finance, R. A. Jarrow, V. Maksimovic, and W. T. Ziemba, eds., vol. 9 of Handbooks in Operations Research and Management Science, Amsterdam, 1995, Institute for Operations Research and the Management Sciences, North-Holland, pp. 31–64.

S. R. PLISKA, *Introduction to Mathematical Finance: Discrete Time Models*, Blackwell Publishers, Malden, Massachusetts, 1997.

R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Englewood, Cliffs, New Jersey, 1972.

R. T. ROCKAFELLAR, *Conjugate Duality and Optimization*, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1974.

M. RUBINSTEIN, *Implied binomial trees*, The Journal of Finanace, 69 (1994), pp. 771–818.

D. SAMPERI, *Inverse Problems, Model Selection and Entropy in Derivative Security Priciong*, PhD thesis, New York University, 1997.

Chapter 28
Volatility Timing and Portfolio Construction Using Realized Volatility for the S&P500 Futures Index

Dimitrios D. Thomakos and Tao Wang

28.1 Introduction

Volatility modeling is important for asset pricing, portfolio choice, option pricing, and risk management. Many studies have built increasingly sophisticated statistical models to capture the characteristics of financial markets' volatility. A lot of earlier work focused on the parametric ARCH and GARCH family of models, on stochastic volatility models, on implied volatility from certain option pricing models, or direct indicators of volatility such as ex-post squared or absolute returns. Partial surveys of the voluminous literature on these models are given by Bollerslev et al. (1994), Ghysels et al. (1996), and Campbell et al. (1997).

However, the trend in recent years have been a renewed interest in obtaining improved daily volatility estimates by using high-frequency, intraday returns to construct daily "realized" or "integrated" volatility. Daily-realized volatility is obtained as the sum of intraday squared returns. Using the theory of quadratic variation, Andersen and Bollerslev (1998) and Barndorff-Nielsen and Shephard (2002) show that the realized volatility estimator, say RV_t, is a consistent estimator of the actual volatility in the sense that $\mathrm{RV}_t \to V_t$, where V_t is the actual (but latent) volatility. This is an interesting result for it is model-free and does not depend upon any particular form of V_t.

Recent work on realized volatility estimation has largely focused on designing estimators with satisfactory statistical properties especially when the data generating process is contaminated from various sources of "microstructure noise." There are many realized volatility estimators that are robust in the presence of such noise and, furthermore, their use is important: if the presence of noise is

D.D. Thomakos (✉)
Department of Economics, University of Peloponnese, Greece
and
Rimini Center for Economic Analysis, Italy
e-mail: thomakos@uop.gr

T. Wang
Department of Economics, Queens College and the Graduate Center of the City University of New York, USA
e-mail: tao.wang@qc.cuny.edu

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_28,
© Springer Science+Business Media, LLC 2010

ignored then many of the earlier volatility estimators are biased and inconsistent. Among these recent estimators that are robust to microstructure noise we have the following: the estimator based on the optimal sampling frequency (Bandi and Russell 2006); the subsampling averaging and the two-time scale estimators (Zhang et al. 2005; Zhang 2006); and the kernel estimators (Zhou 1996; Hansen and Lunde 2006; Barndorff-Nielsen et al. 2006, 2008). There are other volatility estimators that are known to be robust to the presence of discrete jump components in the underlying continuous price process. These include the bipower variation and multipower variation that are particularly useful for detecting jumps. See, for example, Barndorff-Nielsen and Shephard (2004, 2007). Finally, for some recent, relatively complete reviews of the realized volatility literature see Bandi and Russell (2006), Barndorff-Nielsen and Shephard (2007), and McAleer and Medeiro (2008).

A natural question to ask is how these estimators compare in their forecasting ability and, in particular, do they provide a better volatility forecast for representative investors in their portfolio choice? Thus, we evaluate the forecasting ability of different realized volatility estimators from an economic point of view in the context of investment decisions. For our analysis, we consider estimators that are robust in the presence of microstructure noise. To summarize, we consider the following realized volatility estimators:

- The benchmark (Naive) five-minute estimator
- The estimator with the optimal sampling frequency as in Bandi and Russell (2006)
- The subsampling averaging estimator and the two timescale estimators (Zhang et al. 2005)
- The flat-top Bartlett kernel estimator of Barndorff-Nielsen et al. (2006, 2008)

Our analysis builds on the framework developed in Fleming et al. (2001, 2003) and Bandi and Russell (2006). We consider a risk-averse investor who uses conditional mean–variance analysis to allocate funds between a risky asset (here the S&P500) and cash. The investor rebalances his portfolio daily but treats the expected return as constant, since there is very little evidence that expected returns are predictable at the daily level. As a result, the investor follows a volatility-timing strategy to rebalance the portfolio weight for the risky asset by following the change in the one-period ahead conditional variance forecasts. After computing the relevant estimates, we perform a forecasting exercise on them and compare their performance in terms of the "fees" (difference in average utility) that the investor would be willing to pay to use these volatility estimators over the Naive benchmark. Among related studies, where realized volatility estimators are used for economic evaluation, are Fleming et al. (2003) that use the naive realized volatility estimator only but consider four asset classes, and Bandi and Russell (2008) that study volatility forecasting for the purpose of option pricing with a performance comparison from different realized volatility estimators. Our choice of volatility estimators also follows the ones used in the latter paper.

One feature of our study is that we use a very extended data set. Our data spans 14 years of daily returns and realized volatility for the S&P500 futures. Our high-frequency dataset is from the S&P500 futures traded on the Chicago Merchandise

Exchange (from 9:30 a.m. to 16:15 p.m. EST). We also compare the performance of three main estimators, the optimal sampling, the two-time scale estimator, and the kernel estimator, which has not been studied in the volatility timing literature. For forecasting we use two models from the $ARFIMA(p, d, q)$ (autoregressive fractionally integrated moving average) family; the *ARFIMA* models have been heavily used in realized volatility forecasting in the related literature. More technical details are given in Sect. 28.3.

Our results clearly indicate that the estimators are different, both in their forecasting performance evaluation and their economic evaluation, and that the two types of evaluations probably coincide in their results: the estimator matters, the forecasting model matters, the bias correction method matters, and the utility function (and risk aversion) matters. Overall the two timescales estimator of Zhang et al. (2005) appears to be the most robust performer across all estimators we examined.

The rest of the chapter is organized as follows. In Sect. 28.2, we have a brief description of the different realized volatility estimators we use in our study. In Sect. 28.3, we present our data and the two approaches for making "bias corrections" when constructing the realized volatility estimates. In Sect. 28.4, we present our modeling and forecasting methodology as well as the approach for volatility-timing performance measurement criteria. In Sect. 28.5, we evaluate the performance of the volatility-timing strategies and discuss our results. Section 28.6 concludes.

28.2 Realized Volatility Estimators

28.2.1 Preliminaries

This section describes the naive realized volatility estimator as well as the other estimators that have been proposed to reduce bias and improve efficiency under various forms of microstructure noise. To begin with assume that for each day, with time confined in the unit interval [0,1], the observed logarithmic asset prices follow the noise contaminated process:

$$\underbrace{p_{t_i,m} - p_{t_{i-1},m}}_{r_{i,m}} = \underbrace{p^*_{t_i,m} - p^*_{t_{i-1},m}}_{r^*_{i,m}} + \underbrace{u_{t_i,m} - u_{t_{i-1},m}}_{e_{i,m}}, \tag{28.1}$$

where p denotes the observed logarithmic price, p^* denotes the unobservable equilibrium logarithmic price and u denotes the unobservable market microstructure noise. The time index t_i represents the ith observation in the $m+1$ intraday observations with a sampling frequency equal to $1/m$. Assume next that the equilibrium price process evolves as a stochastic volatility local martingale, given as:

$$p^*_{t_i} \overset{\text{def}}{=} \int_0^{t_i} \sigma_s \mathrm{d}W_s, \tag{28.2}$$

where σ_s is a cadlag stochastic volatility process and W_s is standard Brownian motion. The integrated volatility over the whole day is then:

$$V_t \overset{\text{def}}{=} \int_0^1 \sigma_s^2 \mathrm{d}s \tag{28.3}$$

and a consistent estimator for V, as $m \to \infty$, is given by the sum of intraday squared equilibrium returns as:

$$\mathrm{RV}_{t*}^{(m)} \overset{\text{def}}{=} \sum_{i=1}^{m} r_{i,m}^{*2} \to V_t. \tag{28.4}$$

However, $r_{i,m}^{*2}$ is latent and thus the above estimator cannot be implemented. The obvious alternative is to use the sum of intraday squared observable returns but, as discussed in the introduction, this alternative is not robust to the presence of microstructure noise (leading to various inconsistencies). We are, therefore, lead to consider various other estimators. We begin with the naive benchmark, the realized 5-minute estimator.

28.2.2 Naive, 5-Minute Estimator

The prototype realized volatility estimate of ABDL (2001) is simply the sum of the observable intraday squared returns:

$$\mathrm{RV}_t^{(m)} \overset{\text{def}}{=} \sum_{i=1}^{m} r_{i,m}^2. \tag{28.5}$$

In the absence of noise, this estimator is a consistent estimator of V_t as the sampling frequency increases. However, given the existence of microstructure noise, this estimator is inconsistent. The underlying problem can be shown using the expansion:

$$\mathrm{RV}_t^{(m)} = \underbrace{\sum_{i=1}^{m} r_{i,m}^{*2}}_{\mathrm{RV}_*^{(m)}} + \underbrace{\sum_{i=1}^{m} \mathrm{e}_{i,m}^2}_{\mathrm{RV}_u^{(m)}} + \underbrace{2\sum_{i=1}^{m} r_{i,m}^{*}\mathrm{e}_{i,m}}_{2RC_{*,u}^{(m)}} \tag{28.6}$$

and we can see that as $m \to \infty$ we have that $\mathrm{RV}_t^{(m)} \to V_t + m\sigma_{\mathrm{e}}^2 + 2mC$, where $\sigma_{\mathrm{e}}^2 \overset{\text{def}}{=} \mathsf{E}\left[\mathrm{e}_{i,m}^2\right]$ and $C \overset{\text{def}}{=} \mathsf{E}\left[r_{i,m}^{*}\mathrm{e}_{i,m}\right]$. Clearly one cannot recover V but only the sum of the noise variance and the covariance, i.e., $m^{-1}\mathrm{RV}_t^{(m)} \to \sigma_{\mathrm{e}}^2 + 2C$. Typically $\sigma_{\mathrm{e}}^2 \neq 0$ because of the bid-ask bounce. This observation was used by Bandi and Russell (2006) to construct optimal sampling frequency estimators that we discuss below.

To balance the need of using the highest possible sampling frequency and the need of reducing the microstructure noise, ABDL (2001) suggested using 5-minute sparse sampling to estimate V_t in practice. Thus we denote this estimator as:

$$\mathrm{RV}_t^{(m/5)} \overset{\text{def}}{=} \sum_{i=1}^{m/5} r_{i,m}^2. \tag{28.7}$$

28.2.3 Optimally Sampled Estimator

Bandi and Russell (2006) use the properties of $\mathrm{RV}_t^{(m)}$ and determine an optimal sampling frequency $1/m_{\text{opt}}$ based on a mean-squared error (MSE) criterion. They provide exact results as to how this optimal sampling frequency can be chosen under various specifications for C and $\mathrm{e}_{i,m}$ (they include as special cases the conditions $C = 0$ and $\mathrm{e}_{i,m} \sim$ i.i.d in discrete time.) For daily updating the value of m_{opt} is chosen by the following "rule of thumb," valid for large m:

$$m_{\text{opt}} \sim \left(\frac{\widehat{Q}_t}{\widehat{\sigma}_{\mathrm{e}}^4} \right)^{\frac{1}{3}}, \tag{28.8}$$

where the numerator is the estimated daily quarticity, with $Q_t \overset{\text{def}}{=} \int_0^1 \sigma_s^4 \mathrm{d}s$, and the denominator is the square of the estimated noise variance. We have that m_{opt} can be made operational by (a) estimating σ_{e}^2 using the properties of $\mathrm{RV}_t^{(m)}$, that is, using $\widehat{\sigma}_{\mathrm{e}}^2 = m^{-1}\mathrm{RV}_t^{(m)}$ assuming $C = 0$ and (b) estimating the daily quarticity using sparse sampling every, for example, 15-minutes as $\widehat{Q}_t = (m/15) \sum_{i=1}^{m} r_{i,m}^4$. Once m_{opt} is available one can estimate the underlying latent volatility as:

$$\mathrm{RV}_t^{(m_{\text{opt}})} \overset{\text{def}}{=} \sum_{i=1}^{m_{\text{opt}}} r_{i,m}^2. \tag{28.9}$$

Bandi and Russell (2006) provide more details about the practical implementation of this estimator.

28.2.4 Subsample Average and Two Timescales Estimators

In their study, Zhang et al. (2005) categorized $\mathrm{RV}_t^{(m)}$ as the fifth-best realized variance estimator, with a lower sampling frequency estimator, such as $\mathrm{RV}_t^{(m/5)}$, as the fourth-best, and sampling at an MSE-optimal lower sampling frequency as the third-best estimator. The second-best estimator in their categorization is the subsampling estimator.

Suppose the full grid with all observations within the day is defined as G and m is number of observations (the size of G.) G is then partitioned into k nonoverlapping subgrids $G^{(k)}$ of size m_k. If we were to use a sparse sampling (e.g., every 5 or 15-min) approach, we would be using only one portion of the data set. For example, if the highest sampling frequency is every minute and we use 5-min returns to construct a volatility estimator we ignore then other four (4) data points within every 5 min, which potentially could be used. Therefore, one can use additional information while doing sparse sampling. Defining the sparse sampling over subgrid i as:

$$\mathrm{RV}^{(k)}_{\mathrm{sparse}} \stackrel{\mathrm{def}}{=} \sum_{t_j, t_{j+} \in G^{(k)}}^{m_k} (p_{t_{j+}} - p_{t_j})^2, \tag{28.10}$$

where $p_{t_{j+}}$ is the next observation within grid k, we can define the subsample average estimator as the average of all of the possible grids, or subsamples:

$$\mathrm{RV}^{(\mathrm{Avg})}_t \stackrel{\mathrm{def}}{=} \frac{1}{k} \sum_{i=1}^{k} \mathrm{RV}^{(k)}_{\mathrm{sparse}}. \tag{28.11}$$

This estimator, however, is still biased at high frequencies. The first best estimator in Zhang et al. (2005), known as the two timescales estimator, uses $\mathrm{RV}^{(\mathrm{Avg})}$ together with realized volatility calculated at the highest possible frequency possible m, $\mathrm{RV}^{(m)}_t$:

$$\mathrm{RV}^{(\mathrm{TS})}_t \stackrel{\mathrm{def}}{=} \mathrm{RV}^{(\mathrm{Avg})}_t - \frac{\bar{m}}{m} \mathrm{RV}^{(m)}_t, \tag{28.12}$$

where $\bar{m} = (m - k + 1)/k$. The (asymptotically) optimal number of subsamples, k_{opt} can be chosen as:

$$k_{\mathrm{opt}} \stackrel{\mathrm{def}}{=} \left(\frac{3\widehat{\sigma}^4_{\mathrm{e}}}{\widehat{Q}_t} \right)^{\frac{1}{3}} m^{2/3}, \tag{28.13}$$

where $\widehat{\sigma}^2_{\mathrm{e}}$ and $\widehat{Q}_t$ are estimated as in Sect. 28.2.3.

The $\mathrm{RV}^{(\mathrm{TS})}_t$ estimator may be still be improved by selecting the number of subsamples based on finite sample considerations, see Bandi and Russell (2005). It can be shown, see Barndorff-Nielsen et al. (2005, 2007), that $\mathrm{RV}^{(\mathrm{TS})}_t$ can be rewritten as a modified Bartlett-type kernel estimator with the modification term deriving from subsampling. The number of subsamples is now selected as:

$$k^{fs}_{\mathrm{opt}} \stackrel{\mathrm{def}}{=} \left(1.5 \frac{\mathrm{RV}^{(m/15)}_t}{\widehat{Q}_t} \right)^{1/3} m \tag{28.14}$$

and we denote the resulting estimator by $\mathrm{RV}^{(\mathrm{TS,fs})}_t$.

28.2.5 Kernel Estimator

Barndorff-Nielsen et al. (2006, 2008) propose using unbiased (under the assumptions of i.i.d. noise and independence between the noise and equilibrium price processes), flat-top, symmetric kernel estimators of the form:

$$\mathrm{RV}_t^{(\mathrm{kernel})} \stackrel{\mathrm{def}}{=} \widehat{\gamma}_0 + \sum_{j=1}^{q} W_j(\widehat{\gamma}_{-j} + \widehat{\gamma}_j), \tag{28.15}$$

where $\widehat{\gamma}_j \stackrel{\mathrm{def}}{=} \sum_{i=1}^{m-j} r_{i,m} r_{i+j,m}$ is the j-order intraday cross-product and $W_j = K(\frac{j-1}{q})$ is the kernel, with $K(x)$ a function in [0,1] satisfying $K(0) = 1$ and $K(1) = 0$. When $K(x) = 1 - x$ we have the Bartlett kernel, which we use in what follows, and we denote the resulting estimator by $\mathrm{RV}_t^{(\mathrm{Bar})}$. The choice of q, the number of cross-products to use, determines the asymptotic optimality in this class of estimators. It can be shown that for the Bartlett kernel the (asymptotically) optimal number q_{opt} to use coincides with that given by the optimal number of subsamples in the two timescales estimator and we thus have $q_{\mathrm{opt}} \equiv k_{\mathrm{opt}}$.

28.3 Data

In this section, we describe our dataset. The S&P500 futures are traded on the CME. We limit our data to the open cry hours which are from 9:30 EST until 16:15 EST. The S&P500 futures index contract is one of the most heavily traded equity futures contract in the United States. Our data set begins with January 4, 1993 and ends with November 30, 2007. We use the nearby contracts to construct the continuous returns series for the index. To assess the influence of overnight returns (i.e., price changes between 4:15 p.m. and 9:30 a.m. of the following day) when making and using volatility forecasts we use two methods of, so-called, bias correction. The first method multiplies each daily realized volatility estimate RV_t obtained using any of the above methods by a constant factor ζ defined as:

$$\zeta = \frac{\sum_{t=1}^{n} \left(R_t^{\mathrm{S\&P500}}\right)^2}{\sum_{t=1}^{n} \mathrm{RV}_t}, \tag{28.16}$$

where $R_t^{\mathrm{S\&P500}}$ is the daily return on the S&P500 index futures for day t and n is the total number of days in our sample. This procedure guarantees that the average of the transformed variances is equal to the average of the daily squared returns. We denote this first bias correction "without overnights" in the tables. The second method just adds the square of the corresponding overnight returns to each variance estimate and we call it "with overnights" in the tables.

28.4 Methodology

28.4.1 Forecasting Realized Volatility

To be able to assess the economic value of volatility timing, we require out-of-sample forecasts for the various realized volatility estimators that were discussed in the previous sections. To reduce the search for the best possible model for each of these estimators, we utilize two forecasting models that take into account a "stylized fact" about the temporal properties of the realized volatility estimators: long memory. The presence of long memory in realized volatility has been extensively documented in much of the previous literature and we will not attempt to review it here. Letting RV_t denote the observed value of any of the above (bias corrected) realized volatility estimators, it is assumed that it can be modeled as the following (stationary) *ARFIMA*(p, d, q) process, see for example Brockwell and Davis (1991), Beran (1994), and Shumway and Stoffer (1995):

$$(1-L)^d \mathrm{RV}_t \stackrel{\text{def}}{=} X_t, \qquad \phi(L)X_t = \theta(L)\epsilon_t, \tag{28.17}$$

where L is the lag operator, $d \in (-0.5, 0.5)$ is the long memory exponent, $X_t \stackrel{\text{def}}{=} \psi(L)\epsilon_t$ is a causal and invertible ARMA process with $\psi(L) \stackrel{\text{def}}{=} [\phi(L)]^{-1}\,\theta(L)$, and the $\phi(L)$ and $\theta(L)$ are standard (p, q)-order lag polynomials. Finally, ϵ_t is white noise with mean zero and variance σ_ϵ^2. There are $(p+q+2)$ parameters in this type of model which we put in the vector $\gamma \stackrel{\text{def}}{=} (\phi, \boldsymbol{\theta}, d, \sigma_\epsilon^2)^\top$.

For the given range of d the fractional difference operator $(1-L)^d$ is invertible and we can equivalently express RV_t as:

$$\mathrm{RV}_t = (1-L)^{-d}X_t = \sum_{j=0}^{\infty} \tau_j X_{t-j}, \qquad \tau_j \stackrel{\text{def}}{=} \prod_{0<k\le j} \frac{k-1+d}{k} \tag{28.18}$$

with $\tau_0 = 1$.

Suppose that we have available n observations for RV_t. Once appropriate estimates $\widehat{\gamma}_n$ of the parameters γ are available, we can easily use the above representations for computing one-step ahead forecasts of the form:

$$\widehat{RV}_{n+1|n}(\widehat{\gamma}) \equiv \widehat{RV}_{n+1|n} \stackrel{\text{def}}{=} \sum_{j=0}^{n} \widehat{\tau}_j \widehat{X}_{n+1-j|n},\, \widehat{X}_{n+1-j|n} = X_{n+1-j} \text{ for } 1 \le j \le n \tag{28.19}$$

and with $\widehat{X}_{n+1|n}$ being the one-step ahead forecast for the short memory process.

In our application below, and given that our sample sizes are relatively large, we use a two-step estimation procedure in a rolling forecasting context. For computing the forecasts, we split the sample into two parts $n_0 + n_1 = n$ and we use a sliding window of n_0 observations to estimate the parameters γ and then compute and store n_1 one-step ahead forecasts.[1] This way we provide some cushioning from the (unavoidable) changes in the stochastic structure of the realized volatility series over long horizons. The two steps in estimation consists of first estimating the long memory exponent using the semiparametric approach of Robinson (1995), then fractionally differencing the series and estimating the rest of the parameters via conditional least squares. We consider two models in the $ARFIMA(p, d, q)$ class: an autoregressive fractionally integrated model with automatic order selection via the AIC criterion, which we denote by $ARFIMA(AIC, d, 0)$ and by $\mathcal{M}_1$ in the tables, and a more parsimonious $ARFIMA(1, d, 1)$ model which we denote by $\mathcal{M}_2$ in the tables.

28.4.2 Assessing the Statistical and Economic Value of Volatility Forecasts

Once the one-step ahead variance forecasts are available we can assess both their statistical performance and, more importantly for this chapter, their economic value in constructing a dynamic portfolio. It is important that we do both since one anticipates that statistically poor forecasts will not lead to a reasonable economic performance.

To assess the statistical quality of the forecasts, we use two types of postforecasting diagnostic regressions. The first is the well-known forecast unbiasedness regression (" Mincer-Zarnowitz"), where one regresses the actual values on the forecasts in the regression:

$$\mathrm{RV}_{t+1} = \alpha + \beta\widehat{\mathrm{RV}}_{t+1|t} + u_{t+1}. \tag{28.20}$$

Under the hypothesis of forecasting unbiasedness, we have that $\mathsf{E}\left[u_{t+1}|\mathcal{I}_t\right] = 0$, $\alpha = 0$ and $\beta = 1$, which is testable. In addition, the fit of this regression provides a measure of the relative performance of the forecasts in explaining the variability of the dependent variable: higher values indicate a better forecast. With this type of regression we can study the *absolute* performance of each estimator, since we will be comparing the actual values with the forecasts obtained from these values.[2]

A second type of regression, very relevant in the current context, is one that assesses the *relative* forecasting performance of each estimator for any given set of actual values. Here the objective is to see how different forecasts can explain the variability of different realized volatility estimators, which are essentially estimating

[1] We report results for a value of $n_0 = 250$ observations, approximately one trading year. We repeated our analysis using a larger estimation sample of 500 observations but the results were qualitatively similar and available on request.

[2] It is well know that estimating this type of evaluating regression has certain problems, such as biases. These are well documented and are omitted from our discussion.

the same latent quantity. Assuming that we have available a set of $M = 6$ realized volatility estimators and their corresponding forecasts, we estimate the following regression:

$$\mathrm{RV}_{t+1,j} = \beta_{0,j} + \sum_{s=1}^{M} \beta_{s,j} \widehat{\mathrm{RV}}_{t+1|t,s} + \eta_t, \qquad 1 \le j \le M. \tag{28.21}$$

The results from the above regression can potentially help to identify an estimator whose forecasts can explain the variability of not only its own actual values but also that of the actual values of all M volatility estimators. We examine the significance of the estimates of the $\beta_{s,j}$ parameters and the R-squared from this regression. The second type of evaluating regression in (28.19) has been used in Bandi and Russell (2006).

Fleming et al. (2001, 2003) assess the economic value of volatility timing in the context of a mean–variance investor. Consider the return of a risky asset R_{t+1} and a corresponding risk-free rate R^f_{t+1}. We set the latter equal to 6% p.a. in our analysis, following the earlier literature. A simple portfolio with the risky asset and cash has return equal to:

$$R^p_{t+1} \stackrel{\text{def}}{=} R^f_{t+1} + \omega_{t+1|t}\left(R_{t+1} - R^f_{t+1}\right) \tag{28.22}$$

with $\omega_{t+1|t} \in [0, 1]$ being the weight assigned to the risky asset. Assume next that the representative investor has a conditional mean–variance utility function of the form:

$$U_{t+1|t} \stackrel{\text{def}}{=} \mathsf{E}\left[R^p_{t+1}|\mathcal{I}_t\right] - 0.5\lambda \mathrm{RV}^p_{t+1} \tag{28.23}$$

with λ being the parameter of risk-aversion. Using this utility function the optimal weight is given by $\omega_{t+1|t} \stackrel{\text{def}}{=} \mathsf{E}[R_{t+1} - R^f_{t+1}|\mathcal{I}_t]/(\lambda \mathrm{RV}_{t+1})$.

As in Bandi and Russell (2006) we make the above utility and optimal weight operational by substituting the unconditional sample mean of the returns in place of the conditional mean, to abstract from complications dealing with stock return predictability, and the volatility forecasts in place of the unknown latent volatility. Since we can do this for the various volatility estimators, we can compare their performance in constructing a dynamic portfolio based solely on volatility timing. In the obvious notation, the estimated optimal weight is given by:

$$\widehat{\omega}_{t+1|t} \stackrel{\text{def}}{=} \frac{n_1^{-1}\sum_{t=n_0+1}^{n_1}\left(R_t - R^f_t\right)}{\lambda \widehat{\mathrm{RV}}_{t+1|t}} \tag{28.24}$$

while the estimated average (unconditional) utility is given by:

$$\widehat{AU}_{n_1} \stackrel{\text{def}}{=} \widehat{\mu}_{\mathrm{p}} - 0.5\lambda\widehat{\sigma}^2_{\mathrm{p}} \tag{28.25}$$

with $\widehat{\mu}_{\mathrm{p}} \stackrel{\text{def}}{=} n_1^{-1} \sum\limits_{t=n_0+1}^{n} R^{\mathrm{p}}_t$ is the sample mean of the portfolio return over the evaluation period and $\widehat{\sigma}^2_{\mathrm{p}}$ is the correspondingly defined sample variance of the portfolio.

Note that while these quantities are estimated unconditionally, we have that the optimal portfolio weight depends on the volatility forecast and is thus conditionally estimated.

Following the literature, we interpret the difference between average utilities computed on the basis of different volatility estimators as the maximum daily return that the investor would sacrifice to one of the two estimators, the other being the benchmark. That is, we seek the value of the parameter Δ_{ij} such that $\Delta_{ij} \stackrel{\text{def}}{=} \widehat{\text{AU}}_{n_1,i} - \widehat{\text{AU}}_{n_1,j}$ where the ith estimate is compared with the jth estimate who serves as the benchmark.[3] In our analysis, we use the naive 5-minute volatility estimator as our benchmark against which all the other estimators are compared, that is, we report Δ_{i1} since the naive estimator appears first in our results. The values of Δ_{i1} are reported in annualized percentages and when they are positive they indicate that the ith estimator outperforms the benchmark. In addition to calculating Δ_{i1}, we also report the proportion of times that Δ_{i1} is positive in our evaluation sample and we denote it by:

$$P(\Delta_{i1} > 0) \stackrel{\text{def}}{=} \frac{1}{n_1} \sum_{t=n_0+1}^{n_1} I(\Delta_{i1} > 0), \tag{28.26}$$

where $I(\cdot)$ is the indicator function. This will allow us to see the relative frequency with which the competitive estimators outperformed the benchmark.

Finally, to account for the noisy estimation of the utility gains, we apply a recursive scheme in recalculating the average utilities and their respective relative gains across estimators. We repeat the calculation of $\widehat{\text{AU}}$ of (28.25) as well as the calculation of Δ_{i1} but recursively as follows:

$$\widehat{\text{AU}}_\tau \stackrel{\text{def}}{=} \widehat{\mu}_{\text{p},\tau} - 0.5\lambda\widehat{\sigma}^2_{\text{p},\tau}, \quad \Delta_{i1,\tau} \stackrel{\text{def}}{=} \widehat{\text{AU}}_{\tau,i} - \widehat{\text{AU}}_{\tau,1} \quad \text{for} \quad 2 \le \tau \le n_1, \tag{28.27}$$

where the portfolio moments are calculated recursively; for example, the portfolio mean return would be calculated as $\widehat{\mu}_{\text{p},\tau} \stackrel{\text{def}}{=} (n_0 + 1 + \tau)^{-1} \sum_{t=n_0+1}^{n_0+1+\tau} R^{\text{p}}_t$, for $\tau \ge 2$, and similarly for the portfolio variance. At the end of the recursion we can calculate the utility gains at the means of $\widehat{AU}_\tau$, given by $\Delta^{mean}_{i1,rec}$ and the mean utility gains, given by $\bar{\Delta}_{i1,rec}$ which are defined as follows:

$$\Delta^{\text{mean}}_{i1,\text{rec}} \stackrel{\text{def}}{=} \frac{1}{n_1 - 1} \sum_{\tau=2}^{n_1} \left[\widehat{\text{AU}}_{\tau,i} - \widehat{\text{AU}}_{\tau,1}\right], \quad \bar{\Delta}_{i1,\text{rec}} \stackrel{\text{def}}{=} \frac{1}{n_1 - 1} \sum_{\tau=2}^{n_1} \Delta_{i1,\tau}. \tag{28.28}$$

This quantities will help us in better evaluating the economic performance of the various volatility estimators.

[3] Both Fleming et al. (2003) and Bandi and Russell (2006) discuss the issue of noisy estimation of the utility gains from the above procedure. This is unavoidable, although our results are relatively clear about the relative economic gains obtained from specific volatility estimators.

28.5 Performance Evaluation: Results and Discussion

28.5.1 Statistical Performance

Tables 28.1–28.3 present the results on forecasting performance for the various realized volatility estimators. Specifically, Table 28.1 has the results from the forecast unbiasedness regressions of (28.20). Here we are interested in the magnitude and significance of the constant and slope parameters, as well as the relative differences in the goodness of fit of the regressions. A casual look at the table indicates

Table 28.1 Forecast unbiasedness regressions, daily data

	Bias correction without overnights					
	Forecasts from $\mathcal{M}_1$			Forecasts from $\mathcal{M}_2$		
	$\widehat{\alpha}$	$\widehat{\beta}$	R^2_{adj}	$\widehat{\alpha}$	$\widehat{\beta}$	R^2_{adj}
$\text{RV}^{(\text{m}/5)}$	0.000	0.703	0.336	0.000	0.785	0.378
$\text{RV}^{(\text{Avg})}$	0.000	0.871	0.592	0.000	0.909	0.617
$\text{RV}^{(\text{Bar})}$	0.000	0.872	0.592	0.000	0.909	0.617
$\text{RV}^{(\text{TS})}$	0.000	0.871	0.566	0.000	0.957	0.619
$\text{RV}^{(\text{TS,fs})}$	0.000	0.875	0.583	0.000	0.914	0.607
$\text{RV}^{(m_{\text{opt}})}$	0.000	0.851	0.515	0.000	0.896	0.538
	Bias correction with overnights					
	Forecasts from $\mathcal{M}_1$			Forecasts from $\mathcal{M}_2$		
	$\widehat{\alpha}$	$\widehat{\beta}$	R^2_{adj}	$\widehat{\alpha}$	$\widehat{\beta}$	R^2_{adj}
$\text{RV}^{(\text{m}/5)}$	0.000	0.749	0.315	0.000	0.846	0.343
$\text{RV}^{(\text{Avg})}$	0.000	0.787	0.501	0.000	0.840	0.525
$\text{RV}^{(\text{Bar})}$	0.000	0.781	0.501	0.000	0.840	0.528
$\text{RV}^{(\text{TS})}$	0.000	1.135	0.422	0.000	1.366	0.467
$\text{RV}^{(\text{TS,fs})}$	0.000	0.815	0.484	0.000	0.883	0.509
$\text{RV}^{(m_{\text{opt}})}$	0.000	0.822	0.430	0.000	0.919	0.454

Table entries are results on the forecast unbiasedness regressions of (8.20) of the actual, bias-corrected values on the corresponding forecasts. The entries correspond to the estimates of the constant $\widehat{\alpha}$, the slope $\widehat{\beta}$ and R^2_{adj} the adjusted R-squared. The estimators considered are: the naive 5-minute estimator $\text{RV}^{(\text{m}/5)}$, the subsample average estimator $\text{RV}^{(\text{Avg})}$, the flat-top Bartlett estimator $\text{RV}^{(\text{Bar})}$, the two timescales estimators $\text{RV}^{(\text{TS})}$ and $\text{RV}^{(\text{TS,fs})}$, and the optimal sampling estimator $\text{RV}^{(m_{\text{opt}})}$. All calculations are based on one-day ahead variance forecasts from the two models of $ARFIMA(AIC, d, 0)$, $\mathcal{M}_1$, and $ARFIMA(1, d, 1)$, $\mathcal{M}_2$, respectively. The reported estimates are significant at the 1% level of significance. The first forecast is computed using 250 observations, and 3,528 forecasts are used for computing the reported values that the models do a relatively successful job in providing unbiased forecasts but it is also clear that the more sophisticated estimators outperform the naive 5-minute estimator. The constant parameter α is estimated to be zero to three decimals, although it is statistically significant in all cases – possibly a result of the large sample we are using here. The slope parameter β ranges from 0.70 to 0.96, for the models estimated using the "without overnights" bias correction, and from 0.74 to 1.36, for the models estimated using the "with overnights" bias correction. The regression fit is better in the first method of bias correction with a better concentration of the slope estimates around one. For that method using the "without overnights" bias correction, the kernel estimator and the two timescales estimators have R^2 values of around 58%–61%, depending on whether the $\mathcal{M}_1$ or the $\mathcal{M}_2$ model was used in generating the forecasts and we observe that it is estimated closer to one for the more parsimonious $ARFIMA(1, d, 1)$ model and when we use the bias correction method "with overnights."

Estimator	Bias correction	Forecasting model	Rank based on $\widehat{\beta}$
$\mathrm{RV}^{(\mathrm{TS})}$	without overnights	$\mathcal{M}_2$	1
$\mathrm{RV}^{(m_{\mathrm{opt}})}$	with overnights	$\mathcal{M}_2$	2
$\mathrm{RV}^{(\mathrm{TS,fs})}$	without overnights	$\mathcal{M}_2$	3
			Rank based on R^2_{adj}
$\mathrm{RV}^{(\mathrm{TS})}$	without overnights	$\mathcal{M}_2$	1
$\mathrm{RV}^{(\mathrm{Avg})}$	without overnights	$\mathcal{M}_2$	2
$\mathrm{RV}^{(\mathrm{Bar})}$	without overnights	$\mathcal{M}_2$	3

In summary, we can "rank" the estimators based on their estimated slopes $\widehat{\beta}$, their adjusted R-squared across forecasting models and the bias correction methods to see that the overall "winners" are the two timescales estimators $\mathrm{RV}^{(\mathrm{TS})}$ and $\mathrm{RV}^{(\mathrm{TS,fs})}$, the kernel and subsample average estimators $\mathrm{RV}^{(\mathrm{Bar})}$ and $\mathrm{RV}^{(\mathrm{Avg})}$, respectively, and possibly the optimal sampling estimator $\mathrm{RV}^{(m_{\mathrm{opt}})}$. Finally, based on this ranking, the $ARFIMA(1, d, 0)$ model $\mathcal{M}_2$ and the bias correction method "without overnights" appear to be producing the best forecasts. As in Zhang et al. (2005), the two timescales estimator appears to be a top performer.

Turning now to the results from the predictive regressions from (28.21), we note that there are a number of differences based on the bias correction used and the forecasting model used but there is also a clear dominance of the two timescales estimators in terms of significance and magnitude of their coefficient estimates. Looking at Table 28.2 that has the predictive regression results using the "without overnights" bias correction method, and at the top panel where the forecasts are generated using the $\mathcal{M}_1$ model, we can note the following: (a) the subsample average and the kernel estimator do not enter significantly to any of the regressions. The naive 5-minute estimator, on the other hand, does enter significantly in all regressions *except* the regression of the two timescales estimator $\mathrm{RV}^{(\mathrm{TS})}$. Finally, the

Table 28.2 Predictive regressions #1, daily data

	Bias correction without overnights, forecasts from $\mathcal{M}_1$							
	c	$RV^{(m/5)}$	$RV^{(Avg)}$	$RV^{(Bar)}$	$RV^{(TS)}$	$RV^{(TS,fs)}$	$RV^{(m_{opt})}$	R^2_{adj}
$RV^{(m/5)}$	0.000	−0.079	0.119	−0.438	0.063	0.996	0.245	0.417
p^*	0.000	0.083	0.843	0.505	0.116	0.000	0.010	n.a.
$RV^{(Avg)}$	0.000	−0.040	0.420	−0.376	0.149	0.624	0.148	0.604
p^*	0.000	0.219	0.321	0.417	0.000	0.000	0.028	n.a.
$RV^{(Bar)}$	0.000	−0.048	0.307	−0.290	0.142	0.639	0.169	0.604
p^*	0.000	0.135	0.465	0.529	0.000	0.000	0.011	n.a.
$RV^{(TS)}$	0.000	0.016	0.375	−0.110	0.223	0.467	0.005	0.642
p^*	0.274	0.610	0.368	0.809	0.000	0.000	0.940	n.a.
$RV^{(TS,fs)}$	0.000	−0.098	−0.115	−0.063	0.099	0.745	0.331	0.587
p^*	0.000	0.002	0.784	0.891	0.000	0.000	0.000	n.a.
$RV^{(m_{opt})}$	0.000	−0.182	−0.070	−0.214	0.085	0.711	0.557	0.532
p^*	0.000	0.000	0.879	0.670	0.006	0.000	0.000	n.a.
	Bias correction without overnights, forecasts from $\mathcal{M}_2$							
	c	$RV^{(m/5)}$	$RV^{(Avg)}$	$RV^{(Bar)}$	$RV^{(TS)}$	$RV^{(TS,fs)}$	$RV^{(m_{opt})}$	R^2_{adj}
$RV^{(m/5)}$	0.000	−0.067	−0.807	0.046	0.026	1.705	0.062	0.441
p^*	0.147	0.230	0.245	0.952	0.581	0.000	0.585	n.a.
$RV^{(Avg)}$	0.000	0.012	−0.200	0.123	0.184	0.973	−0.127	0.629
p^*	0.058	0.753	0.681	0.822	0.000	0.000	0.112	n.a.
$RV^{(Bar)}$	0.000	0.002	−0.293	0.170	0.172	1.009	−0.101	0.628
p^*	0.025	0.966	0.543	0.753	0.000	0.000	0.204	n.a.
$RV^{(TS)}$	−0.000	0.066	−0.184	0.454	0.306	0.638	−0.264	0.671
p^*	0.421	0.084	0.697	0.392	0.000	0.000	0.001	n.a.
$RV^{(TS,fs)}$	0.000	−0.066	−0.418	−0.097	0.102	1.325	0.098	0.610
p^*	0.002	0.089	0.387	0.858	0.002	0.000	0.215	n.a.
$RV^{(m_{opt})}$	0.000	−0.214	−0.584	−0.128	0.078	1.300	0.490	0.556
p^*	0.003	0.000	0.270	0.830	0.031	0.000	0.000	n.a.

Table entries report results on the predictive regressions of (8.21) of the actual, bias-corrected values on the forecasts from all estimators. The entries are the estimates of $\widehat{\beta}_j$, $j = 0, 1, 2, \ldots, 6$, p^* their p-values and R^2_{adj} the adjusted R-squared. The estimators considered are: the naive 5-minute estimator $RV^{(m/5)}$, the subsample average estimator $RV^{(Avg)}$, the flat-top Bartlett estimator $RV^{(Bar)}$, the two timescales estimators $RV^{(TS)}$ and $RV^{(TS,gs)}$, and the optimal sampling estimator $RV^{(m_{opt})}$. All calculations are based on one-day ahead volatility forecasts from the two models of $ARFIMA(AIC, d, 0)$, $\mathcal{M}_1$, and $ARFIMA(1, d, 1)$, $\mathcal{M}_2$, respectively. The first forecast is computed using 250 observations, and 3,528 forecasts are used for the regressions

two timescales estimators $RV^{(TS)}$ and $RV^{(TS,fs)}$ and the optimal sampling estimator $RV^{(m_{opt})}$ enter in all regressions significantly. The largest estimate in magnitude in all regressions is that of the $RV^{(TS,fs)}$ estimator. All in all, the results from the top panel in Table 28.2 again indicate a superior forecasting and explanatory ability for the two timescales estimators $RV^{(TS)}$ and $RV^{(TS,fs)}$. The results from the lower panel of Table 28.2, where the forecasting model used is $\mathcal{M}_2$, are similar.

Table 28.3 Predictive regressions #2, daily data

Bias correction with overnights, forecasts from $\mathcal{M}_1$

	c	$\mathrm{RV}^{(\mathrm{m}/5)}$	$\mathrm{RV}^{(\mathrm{Avg})}$	$\mathrm{RV}^{(\mathrm{Bar})}$	$\mathrm{RV}^{(\mathrm{TS})}$	$\mathrm{RV}^{(\mathrm{TS,fs})}$	$\mathrm{RV}^{(m_{\mathrm{opt}})}$	R^2_{adj}
$\mathrm{RV}^{(\mathrm{m}/5)}$	0.000	0.185	−0.217	0.054	−0.422	1.429	−0.369	0.343
p^*	0.000	0.015	0.706	0.935	0.000	0.000	0.025	n.a.
$\mathrm{RV}^{(\mathrm{Avg})}$	0.000	0.208	0.755	−0.266	−0.370	0.857	−0.520	0.515
P^*	0.000	0.000	0.074	0.587	0.000	0.000	0.000	n.a.
$\mathrm{RV}^{(Bar)}$	0.000	0.197	0.705	−0.278	−0.376	0.891	−0.478	0.514
p^*	0.000	0.000	0.093	0.567	0.000	0.000	0.000	n.a.
$\mathrm{RV}^{(\mathrm{TS})}$	0.000	0.328	0.255	0.861	−0.304	0.395	−0.856	0.559
p^*	0.000	0.000	0.542	0.075	0.000	0.031	0.000	n.a.
$\mathrm{RV}^{(\mathrm{TS,fs})}$	0.000	0.122	0.461	−0.327	−0.412	0.984	−0.174	0.495
p^*	0.000	0.027	0.272	0.501	0.000	0.000	0.147	n.a.
$\mathrm{RV}^{(m_{\mathrm{opt}})}$	0.000	0.028	0.035	−0.108	−0.393	0.897	0.212	0.442
p^*	0.000	0.638	0.938	0.837	0.000	0.000	0.102	n.a.

Bias correction without overnights, forecasts from $\mathcal{M}_2$

	c	$\mathrm{RV}^{(\mathrm{m}/5)}$	$\mathrm{RV}^{(\mathrm{Avg})}$	$\mathrm{RV}^{(\mathrm{Bar})}$	$\mathrm{RV}^{(\mathrm{TS})}$	$\mathrm{RV}^{(\mathrm{TS,fs})}$	$\mathrm{RV}^{(m_{\mathrm{opt}})}$	R^2_{adj}
$\mathrm{RV}^{(m/5)}$	0.000	0.141	−0.514	0.201	−0.631	1.721	−0.270	0.365
p^*	0.000	0.167	0.035	0.542	0.000	0.000	0.145	n.a.
$\mathrm{RV}^{(\mathrm{Avg})}$	0.000	0.151	−0.207	0.728	−0.485	1.012	−0.546	0.542
p^*	0.000	0.042	0.245	0.002	0.000	0.000	0.000	n.a.
$\mathrm{RV}^{(\mathrm{Bar})}$	0.000	0.142	−0.222	0.670	−0.498	1.067	−0.511	0.541
p^*	0.000	0.053	0.208	0.005	0.000	0.000	0.000	n.a.
$\mathrm{RV}^{(\mathrm{TS})}$	0.000	0.157	−0.068	0.948	−0.409	0.913	−0.843	0.588
p^*	0.011	0.031	0.696	0.000	0.000	0.000	0.000	n.a.
$\mathrm{RV}^{(\mathrm{TS,fs})}$	0.000	0.103	−0.283	0.217	−0.554	1.416	−0.241	0.521
p^*	0.000	0.163	0.109	0.363	0.000	0.000	0.072	n.a.
$\mathrm{RV}^{(m_{\mathrm{opt}})}$	0.000	−0.058	−0.326	0.088	−0.526	1.093	0.439	0.466
p^*	0.000	0.468	0.089	0.733	0.000	0.000	0.002	n.a.

Table entries report results on the predictive regressions of (8.21) of the actual, bias-corrected values on the forecasts from all estimators. The entries are the estimates of $\widehat{\beta}_j$, $j = 0, 1, 2, \ldots, 6$, p^* their p-values and R^2_{adj} the adjusted R-squared. The estimators considered are: the naive 5-minute estimator $\mathrm{RV}^{(\mathrm{m}/5)}$, the subsample average estimator $\mathrm{RV}^{(\mathrm{Avg})}$, the flat-top Bartlett estimator $\mathrm{RV}^{(Bar)}$, the two timescales estimators $\mathrm{RV}^{(\mathrm{TS})}$ and $\mathrm{RV}^{(\mathrm{TS,gs})}$ and the optimal sampling estimator $\mathrm{RV}^{(m_{\mathrm{opt}})}$. All calculations are based on one-day ahead volatility forecasts from the two models of $ARFIMA(AIC, d, 0)$, $\mathcal{M}_1$, and $ARFIMA(1, d, 1)$, $\mathcal{M}_2$, respectively. The first forecast is computed using 250 observations, and 3,528 forecasts are used for the regressions

Finally, the results from Table 28.3, where the bias correction method used was that "with overnights" we see, by and large, similar results to that of Table 28.2 with the following differences: the values of the regression R^2's are generally lower, the coefficient estimates for the $\mathrm{RV}^{(\mathrm{TS,fs})}$ estimator are larger in magnitude than before and the coefficient estimates of the $\mathrm{RV}^{(\mathrm{TS})}$ estimator now enter with a negative sign.

A broad interpretation of our results on the evaluation of the forecasting performance of the various estimators suggests that the more flexible/data-adaptive

$ARFIMA(AIC, d, 0)$ model coupled with the "without overnights" bias correction produced the best overall performance with the top performing estimators being the two timescales estimators $\mathrm{RV}^{(\mathrm{TS})}$ and $\mathrm{RV}^{(\mathrm{TS,fs})}$. In the next section, we examine whether the superior forecasting performance can be translated to superior economic performance as well.

28.5.2 Economic Performance

We next discuss our results for the economic evaluation of volatility forecasts that appear in Tables 28.4–28.7. In all of these tables we report all four evaluation measures discussed in Sect. 28.4.2: the full sample average utility of (28.25) for all competitive estimators, in terms of its difference from the benchmark 5-minute estimator; the proportion of times that the utility of a competitive estimator was higher that the utility of the benchmark estimator from (28.26); the two recursively computed measures of average utility differences from (28.28). We also report results for three values of the risk aversion coefficient $\lambda = 2, 7, 10$ and for both bias correction methods and forecasting models used. We remind the reader that the results on these tables should be interpreted as the annualized fees that an investor would be willing to forego so that he/she could use one of the competitive estimators compared with the benchmark 5-minute estimator.

In Tables 28.4 and 28.5 we present our economic evaluation results when using the "without overnights" bias correction method for the two forecasting models $\mathcal{M}_1$ in Table 28.4 and $\mathcal{M}_2$ in Table 28.5. Looking at the results in Table 28.4 we see that the all estimators outperformed the benchmark 5-minute estimator over 84% of the time during the 3,528 periods of evaluation and that all the evaluation measures indicate positive economic gains. Irrespective of the evaluation measure used we have that the kernel estimator $\mathrm{RV}^{(\mathrm{Bar})}$ produces the more economic gains, closely followed by the two timescales estimators $\mathrm{RV}^{(\mathrm{TS})}$ and $\mathrm{RV}^{(\mathrm{TS,fs})}$. The optimal sampling estimator ranks last in this occasion. As an example, the value of $\Delta_{i1} = 29.76$ for the $\mathrm{RV}^{(\mathrm{TS})}$ estimator means that the percentage difference in mean annualized utility between the naive 5-minute estimator and the $\mathrm{RV}^{(\mathrm{TS})}$ estimator was 29.76%. The shape of the economic evaluation criterion (i.e., the difference in utilities between estimators) can be inferred by looking at the last two columns of Table 28.4 that have the results based on the recursively computed measures. There we can see that the mean difference $\bar{\Delta}_{i1,rec}$ is larger than the difference based on the mean utilities $\Delta^{\mathrm{mean}}_{i1,\mathrm{rec}}$, across values of the risk aversion coefficient, indicating a convex shape. This explains why the full sample differences in average utilities Δ_{i1} in column two fall in between these two recursive measures in columns three and four. Looking at the results in Table 28.5, which correspond to the forecasting model $\mathcal{M}_2$, we see that the economic gains are in general smaller in magnitude except for the case of low risk aversion with $\lambda = 2$, where the $\bar{\Delta}_{i1,\mathrm{rec}}$ measure are higher than in Table 28.4 for all estimators, except for the $\mathrm{RV}^{(\mathrm{Bar})}$ estimator. Here the $\mathrm{RV}^{(\mathrm{TS})}$ estimator is

Table 28.4 Maximum daily returns for using volatility timing, bias correction without overnights, forecasts from $\mathcal{M}_1$

	$P(\Delta_{i1} > 0)$	Δ_{i1}	$\Delta^{\text{mean}}_{1j,\text{rec}}$	$\bar{\Delta}_{1j,\text{rec}}$
		$\lambda = 2$		
$\text{RV}^{(\text{Avg})}$	0.84	17.62	7.94	9.24
$\text{RV}^{(\text{Bar})}$	0.85	37.08	39.35	110.49
$\text{RV}^{(\text{TS})}$	0.84	29.76	10.79	30.33
$\text{RV}^{(\text{TS,fs})}$	0.84	19.78	17.17	36.69
$\text{RV}^{(m_{\text{opt}})}$	0.87	8.98	14.04	22.81
		$\lambda = 7$		
$\text{RV}^{(\text{Avg})}$	0.84	5.51	3.03	10.03
$\text{RV}^{(\text{Bar})}$	0.85	11.60	15.00	23.93
$\text{RV}^{(\text{TS})}$	0.84	9.31	4.11	14.62
$\text{RV}^{(\text{TS,fs})}$	0.84	6.19	6.54	14.63
$\text{RV}^{(m_{\text{opt}})}$	0.87	2.81	5.35	9.45
		$\lambda = 10$		
$\text{RV}^{(\text{Avg})}$	0.84	3.90	2.21	6.20
$\text{RV}^{(\text{Bar})}$	0.85	8.21	10.94	15.83
$\text{RV}^{(\text{TS})}$	0.84	6.59	3.00	9.02
$\text{RV}^{(\text{TS,fs})}$	0.84	4.38	4.77	9.33
$\text{RV}^{(m_{\text{opt}})}$	0.87	1.99	3.90	6.12

Table entries report results on the economic performance evaluation of the volatility estimators when compared with the benchmark 5-minute estimator $\text{RV}^{(m/5)}$. The entries per column report the following: in column one we report the proportion of times during the evaluation period that the competitive estimator had higher utility than the benchmark estimator, see (8.26); in column two we report the difference in average utilities, see (8.25) and what follows, between the competitive estimator and the benchmark estimator; in columns three and four we report the recursive measures of (8.28), namely the difference in average utilities and the average difference in utilities. All entries to columns two to four are expressed in annualized percentages. The competitive estimators considered are: the subsample average estimator $\text{RV}^{(\text{Avg})}$, the flat-top Bartlett estimator $\text{RV}^{(\text{Bar})}$, the two timescales estimators $\text{RV}^{(\text{TS})}$ and $\text{RV}^{(\text{TS,gs})}$, and the optimal sampling estimator $\text{RV}^{(m_{\text{opt}})}$. All calculations are based on one-day ahead volatility forecasts from the two models of $ARFIMA(AIC, d, 0)$, $\mathcal{M}_1$, and $ARFIMA(1, d, 1)$, $\mathcal{M}_2$, respectively. The first forecast is computed using 250 observations, and 3,528 forecasts are used for the regressions

significantly better than the other estimators since an investor would pay almost double the fees to use it when compared with the other competitive estimators.

In Tables 28.6 and 28.7 we report our results when using the "with overnights" bias correction method, again for the two forecasting models we used in generating our forecasts. The results are very similar to those in Table 28.5, in the sense that the $\text{RV}^{(\text{TS})}$ estimator greatly outperforms all other estimators and that the relative

Table 28.5 Maximum daily returns for using volatility timing, bias correction without overnights, forecasts from $\mathcal{M}_2$

	$P(\Delta_{i1} > 0)$	Δ_{i1}	$\Delta_{1j,\mathrm{rec}}^{\mathrm{mean}}$	$\bar{\Delta}_{1j,\mathrm{rec}}$
		$\lambda = 2$		
$\mathrm{RV}^{(\mathrm{Avg})}$	0.84	14.36	5.72	44.68
$\mathrm{RV}^{(\mathrm{Bar})}$	0.84	13.43	5.30	44.66
$\mathrm{RV}^{(\mathrm{TS})}$	0.85	28.58	25.99	82.72
$\mathrm{RV}^{(\mathrm{TS,fs})}$	0.84	10.24	4.42	40.22
$\mathrm{RV}^{(m_{\mathrm{opt}})}$	0.84	3.89	1.60	22.09
		$\lambda = 7$		
$\mathrm{RV}^{(\mathrm{Avg})}$	0.84	4.59	2.27	7.47
$\mathrm{RV}^{(\mathrm{Bar})}$	0.84	4.30	2.10	7.41
$\mathrm{RV}^{(\mathrm{TS})}$	0.85	9.14	10.30	15.21
$\mathrm{RV}^{(\mathrm{TS,fs})}$	0.84	3.28	1.75	6.61
$\mathrm{RV}^{(m_{\mathrm{opt}})}$	0.84	1.25	0.64	3.55
		$\lambda = 10$		
$\mathrm{RV}^{(\mathrm{Avg})}$	0.84	3.26	1.66	4.82
$\mathrm{RV}^{(\mathrm{Bar})}$	0.84	3.05	1.54	4.77
$\mathrm{RV}^{(\mathrm{TS})}$	0.85	6.49	7.56	10.26
$\mathrm{RV}^{(\mathrm{TS,fs})}$	0.84	2.33	1.29	4.25
$\mathrm{RV}^{(m_{\mathrm{opt}})}$	0.84	0.88	0.47	2.26

Table entries report results on the economic performance evaluation of the volatility estimators when compared with the benchmark 5-minute estimator $\mathrm{RV}^{(\mathrm{m}/5)}$. The entries per column report the following: in column one, we report the proportion of times during the evaluation period that the competitive estimator had higher utility than the benchmark estimator, see (8.26); in column two, we report the difference in average utilities, see (8.25) and what follows, between the competitive estimator and the benchmark estimator; in columns three and four, we report the recursive measures of (8.28), namely the difference in average utilities and the average difference in utilities. All entries to the columns from two to four are expressed in annualized percentages. The competitive estimators considered are: the subsample average estimator $\mathrm{RV}^{(\mathrm{Avg})}$, the flat-top Bartlett estimator $\mathrm{RV}^{(\mathrm{Bar})}$, the two timescales estimators $\mathrm{RV}^{(\mathrm{TS})}$ and $\mathrm{RV}^{(\mathrm{TS,gs})}$, and the optimal sampling estimator $\mathrm{RV}^{(m_{\mathrm{opt}})}$. All calculations are based on one-day ahead volatility forecasts from the two models of $ARFIMA(AIC, d, 0)$, $\mathcal{M}_1$, and $ARFIMA(1, d, 1)$, $\mathcal{M}_2$, respectively. The first forecast is computed using 250 observations, and 3,528 forecasts are used for the regressions

economic value of the other estimators compared with the benchmark 5-minute estimator is very small (even slightly negative in certain cases). The results in these tables are important because they indicate that (a) there is a potential advantage in using the first method of bias corrections (the one "without overnights") and that (b) the $\mathrm{RV}^{(\mathrm{TS})}$ estimator appears to be the most robust estimator across forecasting models, bias correction methods, and economic evaluation measures used.

Table 28.6 Maximum daily returns for using volatility timing, bias correction with overnights, forecasts from $\mathcal{M}_1$

	$P(\Delta_{i1} > 0)$	Δ_{i1}	$\Delta^{\text{mean}}_{1j,\text{rec}}$	$\bar{\Delta}_{1j,\text{rec}}$
		$\lambda = 2$		
$RV^{(\text{Avg})}$	0.84	9.06	0.56	12.40
$RV^{(\text{Bar})}$	0.84	6.54	−1.19	10.18
$RV^{(\text{TS})}$	0.96	37.70	35.41	47.04
$RV^{(\text{TS,fs})}$	0.84	3.55	−1.58	7.25
$RV^{(m_{\text{opt}})}$	0.64	−2.70	0.43	0.87
		$\lambda = 7$		
$RV^{(\text{Avg})}$	0.84	2.99	0.23	2.67
$RV^{(\text{Bar})}$	0.84	2.16	−0.48	1.91
$RV^{(\text{TS})}$	0.96	12.45	14.20	14.67
$RV^{(\text{TS,fs})}$	0.84	1.17	−0.63	1.20
$RV^{(m_{\text{opt}})}$	0.64	−0.89	0.17	0.20
		$\lambda = 10$		
$RV^{(\text{Avg})}$	0.84	2.13	0.17	1.69
$RV^{(\text{Bar})}$	0.84	1.54	−0.35	1.14
$RV^{(\text{TS})}$	0.96	8.88	10.45	10.56
$RV^{(\text{TS,fs})}$	0.84	0.84	−0.47	0.68
$RV^{(m_{\text{opt}})}$	0.64	−0.63	0.13	0.14

Table entries report results on the economic performance evaluation of the volatility estimators when compared with the benchmark 5-minute estimator $RV^{(\text{m}/5)}$. The entries per column report the following: in column one, we report the proportion of times during the evaluation period that the competitive estimator had higher utility than the benchmark estimator, see (8.26); in column two, we report the difference in average utilities, see (8.25) and what follows, between the competitive estimator and the benchmark estimator; in columns three and four, we report the recursive measures of (8.28), namely the difference in average utilities and the average difference in utilities. All entries to the columns from two to four are expressed in annualized percentages. The competitive estimators considered are: the subsample average estimator $RV^{(\text{Avg})}$, the flat-top Bartlett estimator $RV^{(\text{Bar})}$, the two timescales estimators $RV^{(\text{TS})}$ and $RV^{(\text{TS,gs})}$, and the optimal sampling estimator $RV^{(m_{\text{opt}})}$. All calculations are based on one-day ahead volatility forecasts from the two models of *ARFIMA*$(AIC, d, 0)$, $\mathcal{M}_1$, and *ARFIMA*$(1, d, 1)$, $\mathcal{M}_2$, respectively. The first forecast is computed using 250 observations, and 3,528 forecasts are used for the regressions

All in all, the results from the economic evaluation of these competing estimators is in agreement with the results on the statistical evaluation on the same estimators. The *ARFIMA*$(AIC, d, 0)$ model $\mathcal{M}_1$ has the best statistical and economic performance when coupled with the first bias correction method, the one "without overnights." The two timescales estimator $RV^{(\text{TS})}$, the kernel estimator $RV^{(\text{Bar})}$ and the subsample average estimator $RV^{(\text{Avg})}$ perform quite well, with the first of these

Table 28.7 Maximum daily returns for using volatility timing, bias correction with overnights, forecasts from $\mathcal{M}_2$

	$P(\Delta_{i1} > 0)$	Δ_{i1}	$\Delta_{1j,\text{rec}}^{\text{mean}}$	$\bar{\Delta}_{1j,\text{rec}}$
		$\lambda = 2$		
$RV^{(\text{Avg})}$	0.84	7.18	5.63	34.67
$RV^{(\text{Bar})}$	0.84	6.12	4.88	33.31
$RV^{(\text{TS})}$	0.86	45.02	40.84	83.46
$RV^{(\text{TS,fs})}$	0.84	3.69	4.57	27.29
$RV^{(m_{\text{opt}})}$	0.56	−1.02	4.89	11.46
		$\lambda = 7$		
$RV^{(\text{Avg})}$	0.84	2.39	2.23	6.68
$RV^{(\text{Bar})}$	0.84	2.04	1.93	6.32
$RV^{(\text{TS})}$	0.86	14.98	16.18	20.17
$RV^{(\text{TS,fs})}$	0.84	1.23	1.81	5.31
$RV^{(m_{\text{opt}})}$	0.56	−0.34	1.94	2.51
		$\lambda = 10$		
$RV^{(\text{Avg})}$	0.84	1.70	1.64	4.33
$RV^{(\text{Bar})}$	0.84	1.45	1.42	4.09
$RV^{(\text{TS})}$	0.86	10.70	11.88	14.02
$RV^{(\text{TS,fs})}$	0.84	0.88	1.33	3.45
$RV^{(m_{\text{opt}})}$	0.56	−0.24	1.42	1.71

Table entries report results on the economic performance evaluation of the volatility estimators when compared with the benchmark 5-minute estimator $RV^{(\text{m}/5)}$. The entries per column report the following: in column one, we report the proportion of times during the evaluation period that the competitive estimator had higher utility than the benchmark estimator, see (8.26); in column two, we report the difference in average utilities, see (8.25) and what follows, between the competitive estimator and the benchmark estimator; in columns three and four, we report the recursive measures of (8.28), namely the difference in average utilities and the average difference in utilities. All entries to the columns from two to four are expressed in annualized percentages. The competitive estimators considered are: the subsample average estimator $RV^{(\text{Avg})}$, the flat-top Bartlett estimator $RV^{(\text{Bar})}$, the two timescales estimators $RV^{(\text{TS})}$ and $RV^{(\text{TS,gs})}$, and the optimal sampling estimator $RV^{(m_{\text{opt}})}$. All calculations are based on one-day ahead volatility forecasts from the two models of *ARFIMA*$(AIC, d, 0)$, $\mathcal{M}_1$, and *ARFIMA*$(1, d, 1)$, $\mathcal{M}_2$, respectively. The first forecast is computed using 250 observations, and 3,528 forecasts are used for the regressions

three being in general the more robust across all cases we examined. Our overall results complement those in the existing literature and clearly show that there is some substantial benefit that can be gained from using these advanced volatility estimators compared with the prototype benchmark estimator.

28.6 Concluding Remarks

In this chapter we present results on the empirical performance of various state-of-the-art estimators of realized volatility using a long dataset on S&P500 futures. Our analysis focus on three aspects: (a) constructing various realized volatility estimators based on intraday returns when the underlying price process is assumed to be contaminated with microstructure noise; (b) computing volatility forecasts from two competing forecasting models and evaluating their statistical performance; and (c) using these forecasts in dynamic portfolio construction and evaluation of the possible utility gains that a representative investor may attain from using these estimators compared with the naive, sparse sampling, 5-minute estimator.

Our results clearly indicate that the estimators are different, both in their forecasting performance evaluation and their economic evaluation, but that the two types of evaluations should probably coincide in their results: the estimator matters, the forecasting model matters, the bias correction matters, and the utility function (and risk aversion) matters. A more flexible and data-adaptive $ARFIMA(AIC, d, 0)$ model has better statistical evaluation for the forecasts it produces, when compared with the more restricted $ARFIMA(1, d, 1)$ model. The overall top performer among the volatility estimators we examined is the two timescales estimator $\mathrm{RV}^{(\mathrm{TS})}$; in some cases its followed in performance by its cousin $\mathrm{RV}^{(\mathrm{TS,fs})}$ with a finite sample-based choice for the subsamples, and in others by either the subsample average estimator $\mathrm{RV}^{(\mathrm{Avg})}$ or the kernel estimator $\mathrm{RV}^{(\mathrm{Bar})}$. The optimal sampling estimator clearly outperforms the naive 5-minute estimator but it does not perform as well as the other estimators in terms of the economic gains it generates.

Our overall results complement those in existing literature in reenforcing the need for carefully constructed volatility estimators that account for the various intraday characteristics of financial markets. A fruitful extension of our work could be the comparison of the performance of the estimators in this study with estimators along the range-based class. This is something we pursue in current research.

References

Andersen, T. G., and T. Bollerslev, (1998), “Answering the skeptics: yes, standard volatility models do provide accurate forecasts,” *International Economic Review* 39, 885–905.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys, (2001), “The distribution of exchange rate volatility,” *Journal of the American Statistical Association*, 96, 42–55.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and H. Ebens, (2001), “The distribution of stock return volatility,” *Journal of Financial Economics*, 61, 43–76.

Bandi, F., and J. R. Russell (2006), “Separating market microstructure noise from volatility,” *Journal of Financial Economics*, 79, 655–692.

Bandi, F., J. R. Russell, and C. Yang (2008), “Realized volatility forecasting and option pricing,” *Journal of Econometrics*, 147, 34–46.

Barndorff-Nielsen, O.E., P. Hansen, A. Lunde, and N. Shephard (2006), “Subsampling realized kernels,” *Journal of Econometrics*, forthcoming.

Barndorff-Nielsen, O.E., P. Hansen, A. Lunde, and N. Shephard (2008), "Designing realized kernels to measure ex-post variation of equity prices in the presence of noise," *Econometrica*, forthcoming, 2008.

Barndorff-Nielsen, O.E., and N. Shephard (2002), "Econometric analysis of realized volatility and its use in estimating stochastic volatility models," *Journal of the Royal Statistical Society B*, 64, 253–280.

Barndorff-Nielsen, O.E., and N. Shephard (2004), "Power and bipower variation with stochastic volatility and jumps (with discussion)," *Journal of Financial Econometrics*, 2, 1–48.

Barndorff-Nielsen, O.E., and N. Shephard (2007), "Variation, jumps and high frequency data in financial econometrics," in R. Blundell, P. Torsten, and W.K. Newey (eds.) Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress, Econometric Society monograph, Cambridge University Press, Cambridge.

Beran, J. (1994), *Statistics for Long Memory Processes*, CRC Press.

Bollerslev, T., R.F. Engle, and D.B. Nelson (1994), "ARCH Models," in R.F. Engle and D. McFadden (eds.), *Handbook of Econometrics*, Vol. IV, 2959–3038. Amsterdam, North-Holland.

Brockwell P.J., and R.A. Davis (1991), *Time Series: Theory and Methods*, Springer-Verlag, New York.

Campbell, J. Y., A.W. Lo, and A.C. MacKinlay (1997), *The Econometrics of Financial Markets*. Princeton University Press, Princeton, New Jersey.

Fleming, J., C. Kirby, and B. Ostdiek (2001), "The economic value of volatility timing," *Journal of Finance*, 56, 329–352.

Fleming, J., C. Kirby, and B. Ostdiek (2003), "The economic value of volatility timing using realized volatility," *Journal of Financial Economics*, 67, 473–509.

Ghysels, E., A. Harvey, and E. Renault (1996), "Stochastic volatility," in G.S. Maddala (eds.), *Handbook of Statistics*, Vol. 14. Amsterdam, North-Holland.

Hansen, P.R., and A. Lunde (2006), "Realized variance and market microstructure noise (with discussion)," *Journal of Business and Economic Statistics*, 24, 127–218.

McAleer, M., and M.C. Medeiro (2008), "Realized volatility: a review," *Econometric Reviews*, 27, 10–45.

Robinson, P.M. (1995), "Gaussian semiparametric estimation of long range dependence," *Annals of Statistic*s, 23, 1630–1661.

Shumway, R.H., and D.S. Stoffer (1995), *Time Series Analysis and Its Applications*, Springer-Verlag, New York.

Zhang, L. (2006), "Efficient estimation of stochastic volatility using noisy observations: a multi-scale approach," *Bernoulli,* 12, 1019–1043.

Zhang, L., P.A. Mykland, and Y. Ait-Sahalia (2005), "A tale of two time scales: determining integrated volatility with noisy high frequency data," *Journal of the American Statistical Association*, 100, 1394–1411.

Zhou, B. (1996), "High frequency data and volatility in foreign-exchange rate," *Journal of Business and Economic Statistics*, 14, 45–52.

Chapter 29
The Application of Modern Portfolio Theory to Real Estate: A Brief Survey

Timothy W. Viezer

> A good mathematical theorem dealing with economic hypotheses was very unlikely to be good economics; and I went more on the rules – (1) use mathematics as a short-hand language, rather than as an engine of inquiry. (2) Keep them till you have done. (3) Translate into English. (4) Then illustrate by examples that are important in real life. (5) Burn the mathematics. (6) If you can't succeed in (4), burn (3).
>
> — Alfred Marshall (1906)

29.1 Introduction

Unlike their colleagues in the stock and bond markets, institutional real estate investors have been slow to use Modern Portfolio Theory (MPT) in their decision-making processes. Surveys by Wiley, (1976), Webb (1984), Louargand (1992), and Worzala and Bajelsmit (1997) have shown that diversification has slowly entered into the lexicon and decision-making processes of institutional real estate investors, but those that used the quantitative methods espoused by MPT were in the minority. To be sure, not all stock and bond managers use MPT to construct or analyze their portfolios, but the real estate practitioners' unwillingness to use these quantitative tools was due to their discomfort with MPTs reliance on data they saw as unrepresentative and MPTs abstraction from the traditional real estate decision-making process, which has been concerned with the details and specifics of "doing the deal."

Despite its shortcomings, MPT has proven to be more than shorthand for the portfolio selection process; it is a powerful "engine of inquiry." Researchers have spent over two decades developing a body of literature that is rich in the breadth of application of MPT to real estate. They have debated its use and shortcomings, attempted to better understand and refine the data inputs, and test improved estimation tools and optimization models. This paper provides a brief survey of the literature that has applied or tested MPT to primarily private, commercial real estate investment. The paper attempts to demonstrate the diversity of MPT's use in real estate rather

T.W. Viezer (✉)
University of Cincinnati, Cincinnati, OH, USA
e-mail: timothy.viezer@uc.edu

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, DOI 10.1007/978-0-387-77439-8_29,
© Springer Science+Business Media, LLC 2010

than to document every study conducted in this field. The paper also elaborates on some of this author's contributions within this literature. Although some references will be made to "indirect" or securitized, publicly traded real estate (real estate investment trusts or REITs), in many instances these vehicles behave like small-cap stocks and do not have the specific problems – illiquid, indivisible ("lumpy"), private transactions – associated with "direct" equity real estate. Although there is an expansive literature on residential real estate, for most residential real estate owners this amounts to a single investment rather than a portfolio, which is the focus of MPT. Since the reader of this volume is already familiar with the concepts employed in MPT, no attempt will be made to repeat them here.

This paper will first outline the basic context of institutional investment in commercial real estate and note how MPT's strategic approach can complement (but never replace) the industry's traditional transaction disciplines. The following section summarizes how MPT has been applied and tested at three levels: (1) the macro decision of adding real estate to a mixed-asset portfolio; (2) the micro decision of diversifying within a real estate portfolio, and (3) diversifying into international real estate. A long (and on-going) debate exists within the real estate investment literature as to whether it is appropriate to use MPT in real estate. The objections noted in this survey include problems with real estate data, the characteristics of the return distribution, and the contention that MPT abstractions ignore important features of the market and decision-making process. Researchers have acknowledged these objections and devised many possible solutions. After reviewing these responses, the paper ends with some reflections on areas for future research in real estate and how some of the lessons learned in real estate may be applied to hedge funds and private equity.

29.2 Why Use MPT in Real Estate?

The slow adoption of MPT is largely the result of the nature of the real estate market. A significant portion of institutionally owned real estate investments are traded in private markets where, by definition, information costs are higher than in "perfect markets." As a result, most investors there have operated on the assumption that the inefficiencies of a private market can be legally exploited by accessing this "inside" information and trading on it through expert intermediaries. The traditional real estate decision process reflects a bias toward the detail and specifics of "doing the deal." Following the passage of the Employment Retirement Income Security Act (ERISA) of 1974, institutional real estate investment grew at an astounding rate, and the focus then was on acquisition not risk management.

In practice, real estate portfolios are generally built one property at a time. Judgment about the strategic goals of the portfolio are made based on the examination of various markets' demographic and economic variables as well as local real estate conditions as suggested by things like vacancies, rents, and new construction. There are several "submarkets" within a metropolitan area/property type market, and it is here that the real estate investment acquisition staff searches for new investments. As such, the acquisition process can be very time-consuming. A survey by Roulac (1981) found that the median time to acquire a single institutional-grade property was 125 person days.

However, the deal-by-deal approach – even if it, in fact, delivered higher risk-adjusted returns (which is debatable) – is not consistent with the notion of portfolio strategy. An aggregation of "best deals" could include a "great deal" on a particular property in a given market, without concern for whether that deal is a good one compared to other properties in different markets. The ultimate result could be a portfolio laden with under-performing assets. The key question is whether a consciously formulated strategy with more attention and resources devoted to policy is better than one that has been arrived at by focusing on transactions. The latter could have been arrived at through trial and error and market feedback regarding what worked and what did not work (at best) or haphazardly (at worst). The answer is not as obvious as it might appear. Some portfolios produced with the latter strategy could have superior returns by way of chance and circumstances. What should distinguish investment managers over the long-run is their ability to repeatedly deliver good portfolio returns. A systematic strategy – such as MPT – should give the manager or analyst a process that can calibrate their judgment, to dissect positive and negative outcomes, and to learn and adjust their strategy from past performance. This is not available to the investment professional who has the lucky coincidence of being in the right place at the right time.

Although the paucity of data restricted research in this area, two lines of research developed over time. In the early 1980s, the focus was on diversifying multiasset portfolios by including real estate. In the latter part of the decade and into the 1990s, researchers turned their attention to diversification "within real estate" portfolios.

29.3 How MPT Is used in Real Estate?

The passage of ERISA made it permissible for pension funds to expand into alternative asset classes such as real estate. ERISAs standards had evolved to the point of applying a "prudent expert" standard to investment decision-making. Pension managers needed to determine whether real estate should be added to their funds of stocks and bonds, and if so, how much of the fund should be allocated to real estate. MPT appeared to rise to the standard of decision-making required by ERISA and similar regulation required at the state level. The data inputs for this analysis came from two sources. First, after being in existence for nearly a decade, the returns of commingled real estate funds (CREFs) in the mid-1980s attracted the attention of researchers who used this data in a number of studies. The National Council of Real Estate Investment Fiduciaries (NCREIF) was a second important source of data. Formed under the leadership of the Frank Russell Company in 1977, NCREIF required that its members report data on their property investments, such as net operating income and appraised values (or transaction prices if their property was purchased or sold). Armed with data and a process, real estate researchers spent over a quarter of a century applying MPT to the macro decision of real estate's role in a mixed-asset portfolio.

How Much Real Estate in a Mixed-Asset Portfolio? Friedman's (1971) study was one of the first attempts to apply MPT to real estate. He suggested that models developed to select common stock portfolios could be adapted to real estate, so long as risk and returns could be quantified. Friedman used data from a sample of fifty properties for the years 1963–1968 and data from a random sample of fifty stocks for the same period. Friedman found that when combined into an efficient portfolio, real estate dominated both the common stock and mixed-asset portfolios for the period under consideration. Eight years later, Findlay et al. (1979) reviewed various methods to achieve an optimal real estate portfolio and concluded that the quadratic model developed by Markowitz was simplistic but adequate. In general, real estate investment returns have been found to have a low correlation with other financial assets' returns.

A large number of papers focused on studying real estate returns. One of the first of these studies was by Robichek et al. (1972), who reported that farmland real estate returns were not significantly correlated with other investment returns. A decade later, Miles and McCue (1982) compared 16 REITs and 18 CREFs for the period 1972–1978. They concluded that these funds could provide sufficient diversification for optimizing portfolios and liquidity not otherwise available through individual property investments. Two years later, Miles and McCue (1984) led the next wave of studies examining real estate returns. Their 1984 study focused on results for a large CREF, confirming the statistical correlation between inflation and real estate returns. The issue of whether or not real estate provided a hedge against inflation would be a favorite area of research for another dozen articles by various authors. Real estate generally was shown to be negatively correlated to stocks and bonds, thus offering a valuable diversification strategy. Later, these studies would be criticized for their use of unreliable and unrepresentative data. Nevertheless, they were based on the best data available at the time and helped promote real estate as a legitimate, core asset class. Miles and McCue also found that commercial real estate returns from various regions and property types exhibited low correlations with bond and stocks. Zerbst and Cambon (1984) noted in their review that all studies reporting correlation coefficients between real estate and common stocks reported a negative correlation.

Over a dozen papers, over almost as many years, studied the issue of how much should be allocated to real estate in an optimal portfolio. These studies used various data to represent real estate over various time periods. It is therefore not surprising that the optimal allocation varied across these studies. What is surprising is how much those optimal allocations differed. Fogler (1984) argued that, based on historical data, a minimum of 15–20% of an investor's portfolio should be allocated to real estate to maximize diversification benefits. Fogler's conclusion was supported by Gold (1986), Irwin and Landa (1987), and Firstenberg et al. (1988). Another close result was Brinson et al. (1986), who found that the optimal weight was 20%. Some authors found that very large allocations to real estate were required.

Ziobrowski and Ziobrowski (1997) concluded that 20–30% in real estate improved the entire efficient frontier, but those gains in efficiency were small. At the extreme, Webb and Rubens (1987, 1986) and Webb et al. (1988) concluded that the optimal allocations were on average 43%, 49–83%, and 66%, respectively. Conversely, smaller allocations – in the range of 3–11% – were recommended by Hartzell (1986). Kallberg et al. (1996) concluded the optimal allocation was 9%, but could be higher if smaller properties were available. The lower correlation of small properties resulted in what these authors termed as a "size effect." Giliberto (1992) concluded that real estate returns had to be 10–12% to justify allocations of 10–15%. Ennis and Burik (1991) used the Capital Asset Pricing Model (CAPM) as the equilibrium asset pricing model, in which portfolio selection occurred. The authors used REIT returns, instead of private real estate returns, to raise the volatility estimates, which the authors believed were biased lower by "appraisal smoothing" (a topic to be discussed later). The higher correlations of REITs were justified in the authors' opinion by the common influence on price multiples and cap rates, future cash flow prospects, and the fact that real estate was a part of corporate assets. Miles et al. (1990) also made adjustments to the data in their analysis. They attempted to control for sample selection bias and used transaction data which they believed was appropriately more volatile than appraisal-based returns. Although these adjustments made real estate's returns more consistent with those of other asset classes, real estate still provided attractive diversification benefits. The high recommended allocations were generally at odds with the actual allocations to real estate in institutional mixed-asset portfolios. This led researchers to question the data and the models. The question of how much should be allocated to real estate continued as researchers began to adjust the data inputs, the model, and the estimators.

Within-Real Estate Diversification: Having established real estate as a legitimate, core asset class, researchers turned their attention to the problem of managing the growing real estate portfolios. In essence and as a practical matter, the separation principle was applied to managing real estate portfolios – first decide the optimal allocation of real estate to a multiasset class portfolio, and then decide how to diversify within the real estate portfolio. However, the question to diversify within a real estate portfolio was made difficult by the industry's unwillingness to publish what it considers a competitive advantage. In an industry that attempts to profit from legal insider information, financial information on *individual* properties is not publicly available. In indices like the NCREIF Property Index (NPI), individual property information is masked by aggregation. An aggregate, (e.g., a metropolitan area property) is not released publicly unless it has at least four properties. Therefore, a balance must be struck between reducing these characteristics into a manageable aggregate for strategy, and including enough characteristics to be meaningful and representative. To an extent every real estate asset is unique: only one building can occupy a particular location at a time. Nonetheless, certain generalizations can be made about the fundamental characteristics of these investments. Researchers would spend two decades exploring which fundamental characteristics provided the best dimensions of diversification (i.e., the manner of aggregating properties with publicly available data).

The earliest studies debated whether property type or geographic region provided the best dimensions by which managers could diversify within the real estate portfolios. Miles and McCue (1982) compared the diversification benefits that could be achieved by using geographic regions vs. property types. Using a period of only 6 years, the authors concluded that diversification by property was more effective than by geographic region. Hartzell et al. (1987) argued that diversification within the real estate portfolio should be done by considering Metropolitan Statistical Area (MSA) growth, property type, and lease maturity instead of traditional property type and location. They extended the length of data and came to the opposite conclusion: regional diversification was more important. Grissom et al. (1987) used data from Houston and Austin for the years 1975–1983 to investigate the benefits of diversification within the real estate portfolio. They found that diversifying across markets and property type reduced unsystematic risk more than across just markets or across just property types. Mueller and Laposa (1995) constructed efficient frontiers for a portfolio of property types using NPI data. They concluded that this strategy enhanced investor returns, but noted that property types had cycles and that predicting these future cycles could be difficult. Jud et al. (2002) used occupancy data which had greater availability than returns. They found that occupancy had a 0.39 correlation to returns. Additionally, they found that retail space markets were relatively segmented because metro areas have their own real estate cycles. As a result, diversifying nationally by metro area was superior to diversification within a region. Like many other papers, these authors found that mean-variance optimal portfolio did not include all available assets. Only 20 of the 58 metro areas entered into efficient portfolios. The low correlation between occupancy and returns was probably due to the fact that occupancy only considers the market for the use of space, which is valued by rents and ignores the market for real estate as an investment asset, valued by discounting the future stream of rents. In studying within real estate diversification, researchers began to study what drove returns.

Grissom et al. (1987) argued that real estate returns varied across geographic regions because of the distribution of industries rather than the arbitrary political demarcations on map. Expanding on this idea, Hartzell (1987) employed eight economically cohesive regions for within real estate diversification. These authors concluded that this "economic diversification" was a more effective means of diversifying the real estate portfolio. Malizia and Simons (1991) also found that economically based diversification strategies were superior because they were based on historical economic relationships rather than political boundaries. Mueller and Ziering (1992) took this research one step further by removing the arbitrary geographic restriction altogether and investigating the economic drivers of individual metropolitan areas as the key determinant of efficient diversification.

In searching for these drivers, Ziering and Hess (1995) included local economic indicators such as the crime rate and cost of living index. Hudson-Wilson (1990) and Hudson-Wilson and Wurtzebach (1994) used statistical methods to determine these associations rather than prespecify them. These two studies employed the K-Means clustering algorithm on "derived market returns." This statistical method forced the "data to speak for themselves," and allowed different property types to be grouped together. For instance, Houston retail and industrial could cluster together, whereas the MSA's office and apartments could be grouped in other clusters. Clustering MSA/property types is one method of achieving economic diversification. Goetzmann and Wachter (1995) applied the clustering algorithm to effective rents for 21 MSA office markets and to vacancy rates for 22 MSAs. The authors' findings supported the idea that there are "families" of cities, but did not compare these aggregates with return data in a mean-variance context. They did, however, present a bootstrapping methodology for investigating the robustness of the clustering algorithm.

Evaluating "Within Real Estate" Diversification Strategies. Firstenberg et al. (1988) demonstrated the diversification benefits from selecting different property types and different geographic areas. Viezer (2000) compared 13 different within real estate diversification strategies along the dimensions of geographic location, property type, and economic region. The article refined previous comparisons by adding controls to the experimental design to determine why one strategy was superior to another.

First, a test data set was used to ensure that no diversification strategy possessed data on an MSA-property type (henceforth, referred to as a "market") that was not available to any other method. Second, a single time period (1991Q1–1997Q2) was used in comparing *across* methods. However, for each diversification strategy, the effects of different time periods on each strategy's inputs and recommendations were examined. Third, diversification strategies that produced a similar number of "within real estate" asset classes were compared to each other. This would separate the effects of the method from the effects of varying the number of dimensions. Viezer used two sources of data from which to aggregate markets. The first was MSA employment by Standard Industrial Classification (SIC). Two economic diversification strategies began with the employment data to try to determine the fundamental economic drivers (and therefore the sources of economic risk for real estate markets). The adaptation of the Isserman and Merrifield (1982) method (henceforth, referred to as the "factor/cluster" method) began by analyzing the industry-level employment data in each MSA using shift–share analysis. The "competitive component" of the shift–share analysis was next used in a factor analysis to reduce the number of industries down to a smaller set of fundamental economic factors. These fundamental factors were then used to cluster the MSAs into similar groups (the "F/C Classification"). Groups of four (FC4), seven (FC7), and eight (FC8) were developed. For the four-, seven-, and eight-economic region indices, all the property types for a particular MSA were aggregated together. However, one variation was considered. A 16-group index was constructed using four-economic regions from the factor/cluster method and segregating these four regions by four property types (F4P).

A second economic diversification method was tested: Mueller and Ziering's (1992) Dominant Employment Category (DEC) method. DEC determines its categories by using t-tests to identify metropolitan area/employment category shares that are significantly different from national averages. The DEC method produced seven categories labeled “energy,” “government,” “high tech,” “manufacturing,” “nonconcentrated,” “travel/tourism,” and “white collar.”

A third economic diversification method started directly with the return data. The “cluster” method grouped market returns using the K-Means clustering algorithm. Although choosing the appropriate number of clusters requires judgment, the numbers of clusters were dictated by NCREIF which reports returns geographically on the basis of MSA, four-regions, and eight-regions. The DEC approach to economic diversification used seven groups. Therefore, clusters of four (C4), seven (C7), and eight (C8) were to be used in computing new return indices. Sixteen clusters (C16) were also considered to help determine whether segregating property types was an important strategy in determining the dimensions of real estate diversification.

The return data itself provided methods for aggregating markets. The markets were grouped on the basis of geography (G4 and G8), property type (PT), and a combination of the two (G4P). Once each of the individual market return series had been classified on the basis of the above assignment schemes, their data were aggregated together to produce new return series. The return series were then used to derive efficient frontiers.

Viezer compared the efficient frontiers constructed with different classification methods and controlled the number of dimensions to determine: (a) whether economic regions were better than geographic regions by considering groupings of four and eight; (b) which economic diversification strategy is the best by comparing groupings of seven; (c) whether property types should be segregated within regions by analyzing groups of 16; and finally (d) whether more asset classes are better than fewer asset classes by comparing groups of 4, 7, 8, and 16. Viezer applied the following hypothesis to each of the thirteen diversification strategies. The mean-variance criterion holds that portfolio A from strategy X is better than (or dominates) portfolio B from strategy Y if $E(R_{\mathrm{A}}) \geq E(R_{\mathrm{B}})$ and $\sigma_{\mathrm{B}} \geq \sigma_{\mathrm{A}}$ and at least one inequality is strict (i.e., ruling out the equality). Therefore, each portfolio along the efficient frontier of a particular diversification strategy will be compared to every portfolio along the efficient frontier of a competing strategy. The efficient frontier is typically graphed in two-dimensional space where the expected return is on the vertical axis and the standard deviation (i.e., risk) is measured along the horizontal axis. Graphically, if efficient frontier X is to the “northwest” (higher expected return and lower risk) of a competing efficient frontier Y, then strategy X dominates strategy Y. Diversification strategy X will be considered superior to diversification strategy Y if X’s efficient frontier above its minimum variance portfolio is everywhere to the northwest of Y’s efficient frontier.

The best 4-dimension strategy was property type, the best 7-dimension strategy was DEC, the best 8-dimension strategy was the factor/cluster, and the best 16-dimension strategy was geographic-property type. Comparing these four best strategies graphically clearly demonstrates the superiority of the 16 dimensions – four property types in four geographic regions strategy Fig. 29.1.

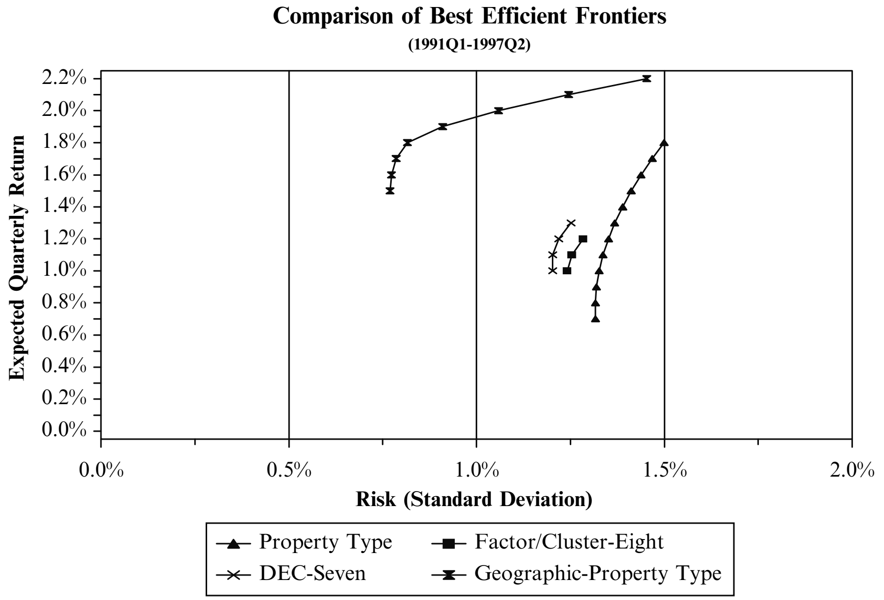


Fig. 29.1 Comparison of best efficient frontiers

He found that more dimensions of diversification were better than fewer, and that property type explained almost one-third of the variation in returns. Viezer also found that the return/risk ratio was more important in explaining the efficiency of a diversification strategy than the associated correlation matrix. The findings had a practical implication: economic diversification was hard to implement as economic diversification group memberships changed over time. This could result in an optimal portfolio in 1 year being inefficient in the next as a market changed cluster membership. The property type/region scheme was far easier to implement.

To discern whether these results may be robust enough to guide actual real estate portfolio management, the reasons for the results should be considered. Regression analysis was used to determine whether the result that property type plays a critical role in optimizing a real estate portfolio. The historical arithmetic average for each market return series was calculated to yield 93 observations. Dummy variables for each of the 13 classification methods were then created and regressed against the return averages. Table 29.1 presents the adjusted R-squares and F-statistics for each of these 13 equations. These statistics show that the cluster method and those methods segregating property types explain real estate investment returns across markets. That the cluster method "explains" as much as 65% of the variation in expected return levels should not be surprising. The cluster method groups markets based on similarities in their return series. The cluster method does not explain why market returns are similar, but only groups those that are similar because the "data themselves say so." Property type does seem to explain around one-third of the variation in market return levels, but adding location to this appears

Table 29.1 Explanatory power of diversification methods

Method	Adjusted R-square	F statistic
Cluster-4	0.4090	22.2195
Factor/cluster-4	−0.0080	0.7562
Geographic-4	−0.0290	0.1365
Property type	0.3322	16.2524
Cluster-7	0.5498	19.7291
DEC	0.0346	1.5507
Factor/cluster-7	0.0541	1.8769
Cluster-8	0.4023	9.8461
Factor/cluster-8	0.0482	1.6648
Geographic-8	0.0022	1.0290
Cluster-16	0.6526	12.5211
Factor/cluster-4 & property type	0.2690	3.2571
Geographic-4 & property type	0.2885	3.4864

to reduce the explanatory power somewhat (note the lower adjusted R-squares for the factor/cluster-property type and geographic region-property type methods).

This low level of explanatory power would seem to support MPT critics like Young and Greig (1993) who contend that the particular circumstances at the individual property level are more important than market level forces. A relatively large property could dominate a return series in markets represented by only a few properties. The F-statistics cited in Table 29.1 test the null hypothesis that expected market returns do not depend on the classification scheme presented by the dummy variables. Consider the case of property type. The critical value of F would lead us to reject the null hypothesis that expected market returns do not depend on property type. The alternative hypothesis is that the expected market returns depend on at least one property type.

The test data set was investigated further using analysis of variance (ANOVA). ANOVA is a statistical technique used to test the equality of three or more sample means and makes inference as to whether the samples come from populations having the same mean. Here, the four samples are property types and the null hypothesis is:

$$\text{Ho} : E(R_{\text{A}}) = E(R_{\text{O}}) = E(R_{\text{R}}) = E(R_{\text{W}}) \tag{29.1}$$

The alternative hypothesis is that the expected returns are not all equal. The F-statistic for the test data set over the 1991Q1–1997Q2 period was 0.8092. Using all periods available for each market, the F-statistic was 0.4179. Based on the critical value for $F_{3,92}$ (2.68), the null hypothesis could not be rejected. This might appear to contradict the case presented thus far for the importance of property type, but it is not entirely damaging. The level of returns is but one factor in the quadratic programming formula that determines the efficient frontier.

In past literature, the presence of low correlations among locations and property types was cited as evidence of potential diversification benefits. Table 29.2 presents a summary of the correlation matrices. Table 29.2 calculates the average and lowest

Table 29.2 Summary of asset class correlations

	Test period			Extended period		
Method	Average correlation (%)	Standard deviation (%)	Minimum correlation (%)	Average correlation (%)	Standard deviation (%)	Minimum correlation (%)
C4	60.50	15.00	33.10	53.70	31.00	−12.70
FC4	88.50	4.20	81.00	54.70	16.00	33.40
G4	89.60	4.60	82.50	56.60	13.50	37.60
PT	80.60	6.90	72.80	56.40	18.60	34.40
C7	50.70	22.00	10.90	51.00	21.30	13.80
FC7	65.00	18.20	12.20	50.60	25.10	2.90
DEC	68.60	19.30	31.70	43.10	27.20	0.70
C8	38.80	24.90	−23.60	38.30	25.50	−25.20
FC8	68.80	18.60	20.40	49.30	26.50	10.60
G8	77.80	12.20	44.00	54.30	16.00	21.10
C16	38.00	24.20	−24.00	38.00	24.20	−24.00
F4P	61.90	17.40	10.90	48.20	22.00	−23.10
G4P	64.00	19.20	13.30	60.00	22.10	−8.40

correlation coefficient of the method's correlation matrix. Table 29.2 also measures the dispersion of these correlation coefficients within a matrix by their standard deviation. These summary measures are presented for the test period and for the full time range available to each method.

For 85% of the possibilities examined, increasing the number of asset classes lowered the average correlation; lowered the minimum correlation; and increased the standard deviation of the correlations. This further supports the notion that more diversification dimensions were better than fewer dimensions. However, the cluster method always produced the lowest average correlation, the smallest individual correlations, and the largest standard deviation of correlations. This ability should have made the cluster method the ideal diversification strategy, but it was not the dominant efficient frontier. Again, expected returns and correlations are not the only variables. Two other important parts of the optimization procedure remain: the portfolio weights and standard deviations within the market return series.

The quadratic programming procedure varies the weights of asset classes to minimize the risk of the portfolio for a given level of expected return. A singular result concerning the optimal portfolio weights is that these weights are dramatically different than those of naive diversification. The optimal asset weights on the efficient frontier typically would require portfolio managers to ignore, on average, two-thirds of the available asset classes and concentrate their bets. Table 29.3 presents the ex post probability that an asset class would receive a zero weighting in the optimal feasible portfolio. Note that for the geographic/property type method, a regional property type would be excluded 77.3% of the time. Moreover, this probability appears to be a positive function of the number of asset classes considered. Both of these findings might appear counter-intuitive. Diversification might colloquially mean lowering risk by spreading investments, that is, by "not putting all

Table 29.3 Probability that an asset class received a zero weighting in the optimal portfolios

Probability that an asset class will receive a zero weighting in the feasible portfolios	Test period (%)	Extended period (%)
Cluster-4	50.0	50.0
Factor/cluster-4	35.7	33.3
Geographic-4	15.0	28.1
Property type	50.0	30.0
Cluster-7	55.7	54.1
DEC	61.9	34.1
Factor/cluster-7	71.4	50.5
Cluster-8	61.9	61.2
Factor/cluster-8	67.7	58.1
Geographic-8	69.2	59.6
Cluster-16	80.6	80.6
Factor/cluster-4 & property type	71.9	69.4
Geographic-4 & property type	77.3	73.2
Weighted average	69.0	63.0

one's eggs in one basket." Yet the results here suggest that not only is this costly to a real estate portfolio manager, it is also inefficient. To spread investments over new markets is often cited as costly to real estate investment professionals given the local nature of management expertise, information costs, etc. No less surprising is the finding that increasing the number of asset classes decreases the percent of asset classes included in the optimal portfolio.

The small number of asset classes included in optimal real estate portfolios has precedents in the literature. Myer and Webb (1991) produced a single optimal portfolio allocation composed of only two property types: retail (42.3%) and industrial (warehouse 50.7% plus office/R&D 6.9%). Mueller and Laposa (1995) noted that their MPT-efficient portfolio mixes did not have all property types included in any given mix. Although these authors did not use the same sample of data nor the same time period, their findings for the 1983Q1–1990Q1 period did correspond to the findings in this study with regard to property type mix. Mueller and Laposa's optimal minimum return portfolio had property type weights: apartment (58%), office (11%), retail (23%), and warehouse (7%). The minimum return optimal portfolio in this study produced weights: apartment (47%), office (8%), retail (24%), and warehouse (21%). The Mueller and Laposa maximum return optimal portfolio assigned 100% weight to retail, whereas the maximum return optimal portfolio in this study allocated 27% to apartments and 73% to retail. In an earlier study by Mueller (1993), the optimal portfolios used only two of eight economic base categories for the time periods 1973Q4–1976Q4 (75% in FIRE and 25% in Government) and 1977Q1–1982Q4 (7.4% in FIRE and 2.6% in Diversified). The period 1983Q1–1990Q4 used only two of the eight Salomon regions (95% in New England and 5% in Southern California). The practical implications of optimal weighting

Table 29.4 Ranking of diversification strategies by return/risk ratios

Method	Average quarter return (%)	Average standard deviation (%)	Average return divided by average standard deviation (%)	Average of asset class return/risk ratios (%)	Weighted average of asset classes return/risk ratios (%)
G4PT	0.96	2.01	48	61	165
FC4PT	1.05	2.05	51	59	72
PT	0.92	1.97	55	55	51
FC4	0.83	1.88	50	50	75
G8	0.76	2.11	36	41	60
FC8	0.76	2.24	34	41	78
G4	0.70	2.07	37	37	58
DEC	0.65	2.09	31	36	78
FEC7	0.79	2.45	32	35	52
C4	1.04	4.59	34	34	72
C7	0.83	4.23	20	29	95
C8	0.71	4.48	16	19	73
C16	0.51	4.03	13	18	54

recommendations are that these weights may be more useful for building new portfolios rather than re-balancing existing portfolios due to the transaction costs.

The choice of which assets to weight depends on their return/risk trade-offs. While the weights were control variables, the correlation, standard deviations, and expected returns were given in the optimization procedure. Correlation matrices were poor indicators of which diversification method would provide the best efficient frontier. The diversification method that produced the lowest correlations (on average or overall) did not provide the best efficient frontier. The return/risk ratio, on the other hand, was pivotal. Table 29.4 presents a summary of the test data's return and risk. Column two is the average expected return across diversification strategies' asset classes. For instance, the average of the four geographic regions' expected returns was 0.70%. The average of the asset classes' standard deviations is given in column three of Table 29.4. Continuing with the example, the average of the four regions' standard deviations was 2.07%. The average return divided by the average standard deviation sometimes differed from the average of the method's return/risk ratios, so both are included. The last column uses the weights of the minimum variance portfolio to derive a weighted average of the return/risk ratio. The average ratio was then rank-ordered. This ranking illustrates the dominance of the geographic/property type diversification strategy. The cluster methods were penalized for the high return/high risk portfolios that produced low return/risk ratios.

The debate of whether property type or region provided greater diversification expanded to include data outside the US Eichholtz et al. (1995) studied correlation matrices, efficient frontiers, and performed principal component analysis of US and UK data. The authors found that the answer was less clear in the UK than in the US Lee (2001) used regression analysis of UK real estate data and found that performance was largely property type driven. The authors found that a portfolio experienced greater tracking error, tilting to property type than to region, and that there was greater diversification across property types in a region, than across regions in a property type. Wellner and Thomas (2004) studied portfolios that were constructed to provide the maximum return, minimum variance, and maximum Sharpe ratio against naïvely diversified (equal weighted) portfolios from ten countries and found that diversification across different countries produced the strongest effects, followed by property type. Geographic diversification within a country trailed the first two methods.

International Diversification: As data became more available in other countries, researchers began to study whether international diversification was effective in real estate. Wilson and Zurbruegg (2003) reviewed over two dozen studies that focused on international diversification of real estate. They found diametrically opposed outcomes which were partly explained by whether the study focused on direct or indirect (securitized) real estate. On balance, the reviewers found that the evidence was more supportive of the proposition that international diversification offered benefits to a portfolio of real estate. Whether this diversification is greater than what could be achieved through other financial instruments was not a settled question. Finally, some authors directed their research in the opposite direction, drilling down into intracity diversification. Pagliari et al. (1995) suggested dividing real estate holdings into urban and suburban areas. Brown et al. (2000) divided Hong Kong into submarkets. These authors found that intracity geographic diversification produced marginally improved portfolio performance, but in some instances naïve diversification was essentially efficient in this context.

29.4 Objections and Challenges

The problems with applying MPT to real estate started with the industry's data. Since private real estate does not trade on an exchange, investors must rely upon appraisals which occur infrequently. This stale pricing was termed "appraisal smoothing" as it was seen to bias the volatility or returns downward. See for example, Geltner (1991, 1993). A biased estimate of volatility might result in higher allocations of real estate to mixed-asset portfolios because the "true" risk was understated. This appeared to be one cause of the high MPT allocations in early studies.

Michael Young is noteworthy among real estate researchers as a persistent and sophisticated critic of the application of MPT to real estate. Starting in 1993, Young and Greig (1993) began by suggesting there was too much emphasis in the literature on constructing portfolios; that individual property analysis was paramount, and that performance depends less on location or property type. The authors felt researchers must investigate property-level factors before uncritically applying MPT to real estate. In a series of studies – King and Young (1994), Young and Graff (1995), Young (2005), Young et al. (2006), and Young (2008) – it was demonstrated that investment risk models with infinite variance provided a better description of individual property returns than normally distributed models. To show how ineffective MPT could be, Young estimated that for an institutional-grade real estate portfolio to achieve 90% risk reduction across multiple risk factors (e.g., location), it would require purchasing most of the institutional grade properties. The authors did note that they were not asserting that MPT is inapplicable in the real estate context but only that the current conceptual version of MPT that has been appropriated without modification from stock market analysis was inapplicable and asset specific risk must be considered in addition to overall market and sector risk. Graff and Young (1996) showed that real estate's covariance/correlation matrices were unstable and their chi-square tests did not show that diversification by property type or property type across metro areas was more efficient than naïve diversification.

There are two conditions (each sufficient but not necessary) that enable an investor to choose a portfolio only on the basis of its expected return and variance. First, if an investor behaves as if they have a quadratic utility function, then they will choose among portfolios on the basis of mean and variance. Second, if returns are normally distributed, then portfolios can be completely described by two parameters: the mean and the variance. Although quadratic utility is a mathematic convenience, some critics challenged MPT first, on the grounds that the distribution of returns was not normally distributed, and second, that the utility function was unrealistic as investors can distinguish between very high and very low returns.

Finally, critics challenged the use of MPT for decision-making with an illiquid asset like real estate. Moreover, most analyses were backward-looking yet investment practice needed forward-looking estimates. As noted by Pagliari et al. (1995), using historical inputs for future behavior may lead to suboptimal results.

29.5 Response to Challenges

Early responses to the challenge of appraisal smoothing were simple but effective. One solution was to use annual data in mixed-asset studies, but this reduced the data available for estimation. In the area of within real estate diversification, researchers ignored the issue altogether as it was assumed to affect all real estate groups equally. A far more sophisticated approach for "unsmoothing" real estate returns was devised by Geltner (1993) and revised by Cho et al. (2003). Newell and MacFarlane (1995) found that the standard risk estimates needed to be increased 80% in mixed-asset decision-making. An improved risk formula by Edelstein and Quan (2006) estimated that appraisal smoothing underestimates both the first and second moments by 50–80%. Correcting mean and variance of real estate returns had levels similar to stocks. Another approach was to use only transaction data. The first study of its kind was Hoag (1980) who used hedonic pricing. Fundamental characteristics were used to estimate industrial property values with regression analysis and this equation was then applied to nontransacting properties. Fisher et al. (2007) devised a transactions-based index using a refined hedonic on NPI data.

The research response to the challenge of nonnormality of the return distribution was at first straightforward, but later became more complex. Schuck (1995) responded to the nonnormality challenge by noting that the returns were indeed leptokurtic, but stated this only makes the optimization process, suggested by MPT, more complicated. Sanders et al. (1995) noted that even if real estate returns have a nonnormal distribution, the realized return would converge on the expected return as the number of repetitions increases. Moreover, stocks appear to have the same problem with nonnormality of returns, yet MPT is widely applied in managing equity portfolios. The nonnormality of returns is important, but statistical procedures appear to be available to deal with this issue. Byrne and Lee (1997) and Lee and Byrne (1998) used the mean absolute deviation (MAD). This process gave less weight to outliers and was seen to be a stable risk substitute for the standard deviation. The (1997) study found that MAD and MPT selected the same assets (regional property types) with only minor variations in their portfolio weights and produced nearly identical efficient frontiers. Byrne and Lee recommended that MPT be applied when the number of assets was less than number of time periods. Maurer and Reiner (2002) used the Lower Partial Mean (LPM) framework to investigate the diversification potential of indirect (securitized) real estate in international investment portfolios containing investments from five countries and analyzed from the point of view of a German and US investor. Their results depended on whether the analysis was ex post or ex ante. There was more significant diversification benefit from ex post (e.g., if the period from which the data inputs were estimated was the same as the portfolio construction period). If the portfolio construction period occurred outside of the estimation period (an ex ante analysis) there were less benefits. International diversification of REITs produced mainly risk reduction rather than return enhancement.

Sivitanides (1998) found that MPT portfolios were only slightly less efficient than the respective downside risk portfolios. A few years later, Sing and Ong (2000) attempted to find a more efficient and robust optimization process by extending the Markowitz model to use LPM and co-LPM. Their proposed downside risk versions of Markowitz's optimization were used to determine the optimal weights for a three asset portfolio in Singapore. The generalized (asymmetric mean-co-LPM) model yielded lower allocations to real estate than the classical Markowitz quadratic programming model. Sing and Ong concluded that the downside risk produced "less risky" portfolios than MPT. However, Cheng and Wolverton (2001) cautioned that comparing classic MPT with downside risk was less straightforward than it might appear. Direct comparison was not appropriate because the two risk measures are different and the exercise was akin to comparing apples to oranges. Cheng (2001) compared the mixed-asset portfolio allocations of classic MPT and downside risk using bootstrapping. Cheng simulated downside risk and MPT return distributions and used Geltner's unsmoothing process to remove appraisal smoothing bias, which is important in the mixed-asset setting. Cheng found that downside risk procedures produced small left tails and larger median returns. Downside risk portfolios also produce more realistic allocations to real estate; 14.4% vs. the MPT allocation in that study of 41.8%. Cheng noted that an investor predisposed to downside risk might choose a different set of assets before optimization. Cheng also found that downside risk is associated with a risk premium and that skewness had significant explanatory power for the cross-sectional variation of property returns.

Not only did downside risk studies address the issue of the nonnormality of return distribution, but they also addressed the question of the investor's proper utility function. Other authors focused on the proper context of an institutional investor noting that most studies ignored an institution's liabilities. Booth (2002) found that efficient real estate allocations were different in an asset–liability framework than those that considered only assets. Similarly, Craft (2001) reported that an asset–liability framework produced allocations to real estate much lower than an asset-only analysis: producing efficient frontier portfolio weights of 0–17.2% vs. 0–41.5%.

Researchers also attempted to propose solutions to the problem of unstable correlations. Lee (2003) found that correlation shifts could be modeled as a function of the market return. The shift coefficients have implications for the skewness of the return distribution. Buetow and Johnson (2001) developed a tactical asset allocation strategy that was fully invested in equity REITs or Treasury bills, depending on the monetary policy environment. The authors found that the correlation structure changes with alternative monetary policy environments and, as a result, the mean-variance optimal allocation varied with monetary policy. Specifically, exposure to REITs should be prominent in a portfolio only in expansive environments.

Many authors noted that ex post analyses produced different recommendations than the more realistic conditions of ex ante analyses. Cheng and Liang (2000) considered the effectiveness of various regional and property type within real estate diversification schemes. They found that the MPT portfolio was statistically more efficient than a naïve portfolio when the test period was the same as the portfolio construction period. However, when the periods were not the same, they found no evidence of statistically significant efficiency. The authors noted that their conclusions were more suggestive than conclusive. That is to say that the lack of statistical significance does not mean there was no benefit for investors who sought every bit of risk reduction, no matter how slight its chance. For those investors, MPT might still be perceived as useful. Some authors believed that the lack of statistical significance could be remedied with better statistical estimators. For example, Bayes–Stein estimators reduce the difference in sample means by effectively shrinking them toward a specified global mean. Efron and Morris (1977) provided the intuition for these estimators. Stevenson (2001) applied the Bayes–Stein shrinkage approach to international real estate securities. The use of the Bayes–Stein estimator led to increased stability in the estimated portfolio allocations and resulted in improved performance. However, the minimum-variance portfolio achieved the greatest improvement in out of sample performance. Stevenson's results held when transaction costs were incorporated. In a later study on US REIT returns, Stevenson (2002) found that unlike the previous study, the minimum variance portfolio underperformed the Bayes–Stein maximum Sharpe ratio portfolio. Stevenson concluded that there was greater performance persistence of sector (property type) than country.

Because of uncertainty, it is possible that any two points – one on the efficient frontier and the other not – could be statistically indistinguishable from each other.

Similar to the work of Michaud (1998), Gold (1995) used bootstrapping to create a "fuzzy" efficient frontier to overcome data quality issues and reduce the need for rebalancing. Gold took property type data from the NPI and recreated the series using 1,000 alternative representations of means and standard deviations through random resampling with replacement. The process left the underlying correlations between assets undisturbed. The fact that two portfolios could be considered statistically indistinguishable would potentially reduce the need for costly trading.

Statistical uncertainty aside, other authors attempted to reshape the MPT tools to be more practitioner-friendly and incorporate real world considerations. Liang and McIntosh (1999) devised a measure that decomposed the overall benefit of an investment into its two components: diversification and return. The measure was scaled in basis points for easy interpretation. Viezer (1999a) devised "diversification hurdle rates" that combined the portfolio allocation process with the opportunistic nature of real estate. Recognizing that it is very costly to trade (or rebalance an existing portfolio of private real estate) and, given that real estate portfolios are built one property at a time, the diversification hurdle rate takes the size of a potential investment, relative to the existing portfolio, and with assumptions about the risk of the new investment solves for the rate of return necessary to increase the current portfolio's Sharpe ratio. The diversification hurdle rate is different for different portfolios and for different sized investments in the same portfolio.

Calculating Diversification Hurdle Rates. The first step is to calculate the portfolio's current simplified Sharpe Ratio. The simplified Sharpe Ratio is the ratio of the portfolio's "expected" (average) return divided by the portfolio's risk as measured by its standard deviation, i.e., $E(r_\mathrm{p})/\sigma_\mathrm{p}$. The expected rate of return on the portfolio $E(r_\mathrm{p})$ with weights w_i in each regional property type (i) is:

$$E(r_\mathrm{p}) = \sum_{i=1} w_i E(r_i)$$

The standard deviation of the portfolio is:

$$\sigma_\mathrm{p} = \left(\sum_{i=1}^{n} w_i^2 \sigma^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \sigma_{i,j} \right)^{1/2} \quad \text{where } i \neq j.$$

The risk-free rate is excluded from this simplified Sharpe Ratio for three reasons. First, as this tool is used for "within real estate" diversification, the risk-free investment is less relevant for the "suboptimization." The risk-free rate is relevant at the "grand strategy" level where the allocation to real estate is decided. Once this is done, the real estate portfolio manager must decide how to invest this amount among choices limited to real estate assets and excluding risk-free investments. Second, this constrains changes in the simplified Sharpe Ratio to factors within the real estate market. Changes in the risk free rate have implications for the real estate allocation rather than choices within real estate. Finally, this simplification scales the returns to levels that are more recognizable by real estate practitioners rather than return differences. Using the portfolio's current weights – that is before any new acquisition is added – the hypothetical portfolio's expected annual return historically over the 1990Q4–1998Q2 period is 4.27% and its standard deviation is 6.95%. Thus, the portfolio's simplified Sharpe ratio is 0.61. The objective is to increase this ratio by adding properties that will increase the portfolio's return per unit of risk. The decision rule is to accept any investment where:

$$\text{New Sharpe Ratio} \geq \text{Old Sharpe Ratio}$$

The second step is to recalculate the Sharpe ratio for each regional property type assuming an additional investment in the regional property type. In this example, the proposed investment is \$20 million. The standard deviation is recalculated with the resulting new weights. This assumes that the new investment will have the same risk (standard deviation) as the historical series. The expected return is the weighted average of the 16 regional property types plus the proposed investment. The return on the new investment that would leave the Sharpe Ratio unchanged can be solved for using the decision rule above. The hurdle rate is equal to:

$$\frac{((\text{Old Sharpe Ratio} \times \text{New Portfolio Standard Deviation}) - (\text{Sum of New Weights} \times \text{Old Returns}))}{\text{Weight of New Investment}}$$

The individual diversification hurdle rates in the table below were calculated using historical NCREIF data (1990Q4–1998Q2). The real estate investment community currently uses a variety of benchmarks for determining whether or not to purchase a particular property including cash flow, price per square foot, internal rate of return, and the particular characteristics that make the property a unique asset. The diversification hurdle rates are best used and understood in the context of an investor's minimum acceptable rate of return Table 29.5.

Using Diversification Hurdle Rates. The proper use of the diversification hurdle rates requires a two-step process. A hypothetical investor managing a \$2 billion real estate portfolio who has derived required rates of return for the various regional property types from the CAPM. He has assumed a 7% risk-free rate and a 300 b.p. market premium to which he has multiplied the regional property type's beta. Column four of the Table 29.5presents these required rates of return (RROR). Now suppose the portfolio manager is considering investing \$20 million in a particular East apartment. If the risk of the particular property is assumed to be the same as the larger East apartment subclass, then the incremental return could be calculated to hold the Sharpe ratio unchanged. Suppose that the proposed acquisition first clears the investor's required rate of return of 9.49%. A second comparison indicates that the property would clearly earn better than 6.29% (the diversification hurdle rate) on an annual basis and the portfolio would benefit from its inclusion from a risk-adjusted perspective. Notice that these two hurdles are consecutive but not additive. Conversely, a West office that could first clear the required rate return of 11.56% may (or may not) be able to clear the second hurdle rate of 12.64% (the

Table 29.5 Hurdle rates

Region	Type	Diversification (%)	RROR (%)
East	Apartment	6.29	9.49
East	Office	11.95	11.20
East	Retail	3.74	10.08
East	Warehouse	9.41	9.81
Midwest	Apartment	5.61	9.04
Midwest	Office	10.28	10.80
Midwest	Retail	2.99	9.36
Midwest	Warehouse	6.24	8.67
South	Apartment	3.75	7.88
South	Office	12.50	10.37
South	Retail	3.38	9.28
South	Warehouse	7.42	8.84
West	Apartment	7.68	9.28
West	Office	12.64	11.56
West	Retail	5.75	9.53
West	Warehouse	11.25	10.50

diversification hurdle rate). Consider another example. The diversification hurdle rates in the table imply that the portfolio would benefit from the acquisition of a Midwest retail property that yields at least 2.99% from a risk-adjusted return perspective. However, the required rate of return ensures that no retail property from the Midwest would be acquired unless it yielded at least 9.36%. The greater of the required rate of return and the diversification hurdle rate is the actual litmus test for the decision to acquire a property. The portfolio manager can not only evaluate individual deals with these diversification hurdle rates, but also construct an acquisition strategy. Lower diversification rates suggest regional property types whose weights should be increased in the current portfolio. Looking at the table, the diversification hurdle rates suggest that Midwest retail should be an acquisition target. However, if there are no Midwest retail deals on the market but many West offices that meet the 11.56% required rate of return, then they should be acquired if their expected return exceeds 12.64%.

REITs. Seiler et al. (2001) studied whether REITs could be used to rebalance within a real estate portfolio. The authors compared the property type allocations along the efficient frontier for a private real estate portfolio. Substituting public real estate property type returns for the private real estate property type returns produced different optimal allocations. Although in practice REITs were used to bring a real estate allocation up to target in a mixed-asset portfolio, unfortunately in a mean-variance, within real estate context the two offered a low degree of substitutability.

Data availability had both shaped and limited the research agenda in real estate. For example, the within real estate diversification issue was a response to data availability. To expand the amount of available return data, Viezer (1999c) demonstrated that synthetic real estate returns could be created based on market fundamentals such as rents and value. These "implied market returns" (IMRs) were then compared to NPI returns. Viezer found that at the level of MSA property types, the correlations between the NPI and IMR series were not statistically significant. This was mainly due to membership issues in the NPI data: a large single property could distort an entire MSA property type. But when the IMRs were aggregated, (at the national, regional, and property type levels) the two return series were highly and positively correlated and statistically significant. IMRs had the advantage that they could create return series for areas where no NPI returns existed and the IMRs could also be forecasted. Authors continued to expand analyses with whatever data was available. For example, Shaoqun and Ying (2004) used price and rent data to study real estate diversification for 35 large- and medium-scale cities in China.

The ability to forecast detailed real estate returns could assist both mixed-asset and within real estate diversification studies. As demonstrated in the studies that compared ex ante and ex post recommendations, the use of historical data has its limits for recommending future actions. Patrick Henry was right when he said "I know of no way to judge the future but by the past," but using historical data as inputs for MPT analyses that prescribe future allocations has been likened by critics to driving a car by looking in the rearview mirror. Fortunately, the literature has developed models for forecasting real estate returns. DiPasquale and Wheaton (1992) posited a conceptual framework to explain how real estate's space and capital markets are integrated. Two variables link the space market (where use decisions are made) and capital markets (where investment decisions are made). Rents are determined in space market, but are central to determining the demand for real estate as an investment, as an investor acquires the discounted present value of the future expected income stream. Construction increases supply of investment assets, which lowers prices in asset market as well as rents in the space market. The authors provided equations to graphical framework, but did not estimate these equations. Viezer (1998, 1999b) developed a pooled recursive system that integrated space and capital markets to predict occupancy, real rents, capitalization rates, market value per square foot, net change in stock, and real construction costs. This real estate econometric forecast model (REEFM) also produced synthetic investment returns, called implied market returns, which could forecast over 200 MSA-property type market returns.

29.6 Where to From Here?

G.E.P. Box once said "all models are wrong, but some are useful." Without adaptations, MPT may be "wrong," but it is useful. Developed over half-a-century ago, it has become a standard for comparison, as an engine of inquiry. In comparing portfolio recommendations, it offers richer insights than naïve diversification. MPT has been used to evaluate the inclusion of real estate in a mixed-asset portfolio, as well as how to best diversify within a real estate portfolio, and whether international diversification reduces risk or enhances returns. Nonetheless, the debate over its application to real estate is fresh and continuing. Critics have pointed out that the nonnormality of the return distribution and the instability of correlations in historical data ensures that the use of MPT with this data will be a case of "garbage in, garbage out." They conclude that portfolio construction is futile and all that really matters is the quality of the individual deals.

The critics' facts are beyond dispute, but their conclusion misses two important points. First, MPT calls for expectations of risk and return. Early studies used historical data to understand past relationships and glean some ideas about the future. In this second use there were the greatest problems. But it is no different for investors evaluating individual real estate deals. They must make estimates of future rents, tenant occupancy, inflation, and interest rates. The future is uncertain and perilous for anyone making predictions. No one would recommend abandoning financial analysis due to the problems of forecasting. MPT is a complement that can never replace individual deal analysis. For all of its mechanical precision, the use of MPT requires the art of judgment.

Based on this brief survey, several areas stand out for future research. Lee (2001) showed importance of tracking error. Much of the real estate MPT analysis focused on the absolute return and risk of a portfolio, yet most managers are evaluated against a performance benchmark. Further work on minimizing the tracking error of real estate portfolios appears to be in order. Although the greatest impacts have come from the error of estimating return means, further work on stabilizing the covariance matrix should be done. Researchers might consider the work of Ledoit (1996, 1999), who applied shrinkage to covariance or MacKinlay and Pastor (2000), who found that, when a risk factor is missing from an asset pricing model, the mispricing is embedded within the residual covariance matrix. They exploit this by linking the vector of returns on other assets to the covariance matrix with identity matrix and found that this new covariance matrix performed well in and out of sample. The return distributions of synthetic returns should be tested and the methodology applied to those countries lacking survey-based returns. Researchers might consider applying downside risk and fuzzy efficient frontiers to diversification hurdle rates. Many fund managers do not own individual real estate properties but invest in funds. Olaleye and Aluko (2007) studied real estate diversification with Nigerian managers. Their evidence suggested that manager and property type diversification improved performance. Although their data was limited, this is an important avenue of research for practitioners.

The lessons learned in real estate have been applied to hedge funds. Lhabitant (2006) noted that certain hedge fund strategies have valuation biases because they are not actively traded (merger arbitrage, distressed debt, convertible arbitrage, and emerging markets strategies). He noted that methods were available from real estate researchers that could "unsmooth" these returns. Certainly the last frontier for MPT is private equity.

Like real estate in the 1980s, private equity allocations are growing among institutional investors. Grabenwarter and Weidig (2005) reviewed the nascent field of quantitative risk management in private equity. They acknowledge that quantitative risk management is not easily adaptable to private equity, which is a people's business. Like private real estate, there are no market prices or liquidity. The lack of transparency and free unbiased historical data is also a major impediment to academic research. Thomson Venture Economics, Cambridge Associates, and Wilshire Associates provide return series, but there are issues with comparing internal rates of return to public time-weighted returns for research on the allocation to and within private equity. Perhaps like their private real estate colleagues, private equity investors may someday find MPT as a useful engine of inquiry.

References

Addae-Dapaah, K., Wee, S.G., Ebrahim, M., "Real Estate Portfolio Diversification by Sources of Return." *Journal of Real Estate Portfolio Management* 8 (2002):1–15

Bajtelsmit, V.L., Worzala, E.M., "Real Estate Allocation in Pension Fund Portfolios." *Journal of Real Estate Portfolio Management* 1 (1995):25–38

Benjamin, J.D. Sirmans, G.S., Zietz, E.N., "Returns and Risk on Real Estate and Other Investments: More Evidence." *Journal of Real Estate Portfolio Management* 7 (2001):183–214

Bond, S.A., Patel, K., "The Conditional Distribution of Real Estate Returns: Are Higher Moments Time Varying?" *Journal of Real Estate Finance and Economics* 26 (2003):319–339

Booth, P.M., "Real Estate Investment in an Asset/Liability Modeling Context." *Journal of Real Estate Portfolio Management* 8 (2002):183–198

Brinson, G.P., Diermeier, J.J., Schlarbaum, G.G., "A Composite Portfolio Benchmark for Pension Plans." *Financial Analysts Journal* 42 (1986):15–24

Brown, R.J., Li, L.H., Lusht, K., "A Note on Intracity Geographic Diversification of Real Estate Portfolios: Evidence From Hong Kong." *Journal of Real Estate Portfolio Management* 6 (2000):131–140

Buetow, Gerald W. Jr., Johnson, R.R., "The Real Estate Asset Allocation Decision: Monetary Policy Implications." *Journal of Real Estate Portfolio Management* 7 (2001):215–223

Byrne, P., Lee, S., "Real Estate Portfolio Analysis Under Conditions of Non-Normality: The Case of NCREIF." *Journal of Real Estate Portfolio Management* 3 (1997):37–46

Cambridge Associates LLC. "U.S. Private Equity Index and Selected Benchmark Statistics." June 30 (2007a)

Cambridge Associates LLC. "U.S. Venture Capital Index and Selected Benchmark Statistics." June 30 (2007b)

Center for International Securities and Derivatives Markets. "The Benefits of Real Estate Investment: 2006 Update." Amherst, Massachusetts, University of Massachusetts, (May 2006)

Cheng, P., "Asymmetric Risk Measures and Real Estate Returns." *Journal of Real Estate Finance and Economics* 30 (February 2005):89–102

Cheng, P. "Comparing Downside-Risk and Mean-Variance Analysis Using Bootstrap Simulation." *Journal of Real Estate Portfolio Management* 7 (2001):225–238

Cheng, P., Liang, Y., "Optimal Diversification: Is It really Worthwhile?" *Journal of Real Estate Portfolio Management* 6 (2000):7–16

Cheng, P., Wolverton, M.L., "MPT and the Downside Risk Framework: A Comment on Two Recent Studies." *Journal of Real Estate Portfolio Management* 7 (2001):125–131

Cho, H., Kawaguchi, Y., Shilling, J.D., "Unsmoothing Commercial Property Returns: A Revision to Fisher-Geltner-Webb's Unsmoothing Methodology." *Journal of Real Estate Finance and Economics* 27 (November 2003):393–405

Conover, C.M., Friday, H.S., Sirmans, G.S., "Diversification Benefits from Foreign Real Estate Investments." *Journal of Real Estate Portfolio Management* 8 (2002):17–25

Cole, R., Guilkey, D., Miles, M., Webb, B., "More Scientific Diversification Strategies for Commercial Real Estate." *Real Estate Review* 19 (Spring 1989):59–66

Cooperman, L.G., Einhorn, S.G., Melnikoff, M., "Property Revisited: The Role of Property in Pension Fund Investments." New York, NY: Goldman Sachs, 1984

Corgel, J.B., de Roos, J.A., "Recovery of Real Estate Returns for Portfolio Allocation." *Journal of Real Estate Finance and Economics* 18 (May 1999):279–296

Craft, T.M., "The Role of Private and Public Real Estate in Pension Plan Portfolio Allocation Choices." *Journal of Real Estate Portfolio Management* 7 (2001):17–23

DiPasquale, D., Wheaton, W.C., "The Markets for Real Estate Assets and Space: A Conceptual Framework." *Journal of the American Real Estate and Urban Economics Association* 20 (Summer 1992):181–197

Edelstein, R.H., Quan, D.C., "How Does Appraisal Smoothing Bias Real Estate Returns Measurement?" *Journal of Real Estate Finance and Economics* 32 (February 2006):41–60

Efron, B., Morris, C., "Stein's Paradox in Statistics." *Scientific American* 236 (May 1977):119–127

Eichholtz, P.M.A., Hoesli, M., MacGregor, B.D., Nanthakumaran, N., "Real Estate Portfolio Diversification By Property Type and Region." *Journal of Property Finance* 6 (1995):39–59

Ennis, R.M., Burik, P., "Pension Fund Real Estate Investment Under a Simple Equilibrium Pricing Model." *Financial Analysts Journal* 47 (May–June 1991):20–30

Findlay, M.C., Hamilton, CW., Messner, S.D., Yorkman, J.S., "Optimal Real Estate Portfolios." *Journal of the American Real Estate and Urban Economics Association* 7 (Fall 1979):298–317

Firstenberg, P.M., Ross, S.A., Zisler, R.C., "Real Estate: The Whole Story." *Journal of Portfolio Management* 14 (Spring 1988):22–34

Fisher, J., Geltner, D., Pollakowski, H., "A Quarterly Transactions-Based Index of Institutional Real Estate Investment Performance and Movements in Supply and Demand." *Journal of Real Estate Finance and Economics* 34 (January 2007):5–33

Fisher, J.D., Geltner, D.M., Webb, R.B., "Value Indices of Commercial Real Estate: A Comparison of Index Construction Methods." *Journal of Real Estate Finance and Economics* 9 (1994):137–164

Fisher, J.D., Liang, Y., "Is Property Type Diversification More Important Than Regional Diversifcation?" *Real Estate Finance* 17 (2000):35–40

Fogler, H.R., "Twenty Percent in Real Estate: Can Theory Justify It?" *Journal of Portfolio Management* 10 (1984):6–13

Foresti, S.J., Toth, T.E., "Private Equity Investing Part 2 – Generating Asset Class Assumptions." Wilshire Associates Incorporated. January 23, 2006

Friedman, H.C., "Real Estate Investment and Portfolio Theory." *Journal of Financial and Quantitative Analysis* 6 (March 1971):861–874

Geltner, D.M., "Smoothing in Appraisal-Based Returns." *Journal of Real Estate Finance and Economics* 4 (September 1991):327–345

Geltner, D., "Estimating Market Values From Appriased Values Without Assuming an Efficient Market." *Journal of Real Estate Research* 8 (Summer 1993):325–345

Geltner, D., Goetzmann, W., "Two Decades of Commercial Property Returns: A Repeated-Measures Regression-Based Version of the NCREIF Index." *Journal of Real Estate Finance and Economics* 21 (July 2000):5–21

Geltner, D., Miller, N.G., *Commercial Real Estate Analysis and Investments*. Upper Saddle River, New Jersey: Prentice Hall, 2001

Giliberto, M., "Real Estate Risk and Return: 1991 Survey Results." New York, NY: Salomon Brothers, 1992

Giliberto, M., Hamelink, F., Hoesli, M., MacGregor, B., "Optimal Diversification Within Mixed-Asset Portfolios Using a Conditional Heteroskedasticity Approach: Evidence From the U.S. and the U.K." *Journal of Real Estate Portfolio Management* 5 (1999):31–45

Goetzmann, W.N., Wachter, S.M., "Clustering Methods for Real Estate Portfolios." *Real Estate Economics* 23 (1995):271–310

Gold, R.B., "Real Estate: Can Institutional Portfolios Be Diversified Without It?" Chicago, IL: JMB Institutional Realty Corporation, 1986

Gold, R.B., "Why the Efficient Frontier for Real Estate is 'Fuzzy.'" *Journal of Real Estate Portfolio Management* 1 (1995):59–66

Grabenwarter, U., Weidig, T., *Exposed to the J-Curve: Understanding and Managing Private Equity Fund Investments*. London, England: Euromoney Books, 2005

Graff, R.A., Young, M.S., "Real Estate Return Correlations: Real-World Limitations on Relationships Inferred from NCREIF Data." *Journal of Real Estate Finance and Economics* 13 (September 1996):121–142

Grissom, T.V., Hartzell, D., Liu, C., "An Approach to Industrial Real Estate Market Segmentation and Valuation Using the Arbitrage Pricing Paradigm." *Journal of the American Real Estate and Urban Economics Association* 15 (1987):199–219

Grissom, T.V., Kuhle, J.L., Walther, C.H., "Diversification Works in Real Estate, Too." *Journal of Portfolio Management* (Winter 1987):66–71

Hartzell, D., "Real Estate in the Portfolio." New York, NY: Salomon Brothers, 1986

Hartzell, D., Hekman, J., Miles, M., "Diversification Categories in Investment Real Estate." *Journal of the American Real Estate and Urban Economics Association* 13 (Spring 1985):32–47

Hartzell, D., Shulman, D., Wurtzebach, C., "Refining the Analysis of Regional Diversification for Income Producing Real Estate." *Journal of Real Estate Research* 2 (1987):85–95

Hess, R., Liang, Y., "A Sector View of Public Market Ownership in the United States." *Journal of Real Estate Portfolio Management* 8 (2002):271–284

Higgs, H., Worthington, A.C., "The Prospects for Geographic Diversification in U.K. Regional Property Investment Implications Derived From Multivariate Cointegration Analysis." In: Worthington, A., (ed) *Discussion Paper in Economics, Finance and International Competitiveness*. Queensland University of Technology, Australia, 2002, pp 1–15

Hoag, J.W., "Towards Indices of Real Estate Value and Return." *Journal of Finance* 35 (May 1980):569–580

Hodgson, D.J., Slade, B.A., Vorkink, K.P., "Constructing Commercial Indices: A Semiparametric Adaptive Estimator Approach." *Journal of Real Estate Finance and Economics* 32 (March 2006):151–168

Hudson-Wilson, S., "New Trends in Portfolio Theory." *Journal of Property Management* 55 (May/June 1990):57–58

Hudson-Wilson, S., Wurtzebach, C.H., (ed) *Managing Real Estate Portfolios*. Burr Ridge, IL: Richard D. Irwin, Inc., 1994

Ilkiw, J., "Asset Allocation and Real Estate: How to Make Prudent and Profitable Decisions." *PREA Quarterly* (Summer 2001):2–9

Irwin, S.H., Landa, D. "Real Estate, Futures, and Gold as Portfolio Assets." *Journal of Portfolio Management* 13 (1987):29–34

Isserman, A.M., Merrifield, J., "The Use of Control Groups in Evaluating Economic Policy." *Regional Science and Urban Economics* 12 (February 1982):43–58

Jorion, P., "Bayes-Stein Estimation for Portfolio Analysis." *Journal of Financial and Quantitative Analysis* 21 (September 1986):279–292

Jud, G.D., Wingler, Tony R., Winkler, D.T., "The Integration of Retail Space Markets." *Journal of Real Estate Portfolio Management* 8 (2002):79–92

Kallberg, J.G., Liu, C.H., Greig, D.W., "The Role of Real Estate in the Portfolio Allocation Process." *Real Estate Economics* 24 (1996):359–377

King, D.A. Jr., Young, M.S., "Why Diversification Doesn't Work." *Real Estate Review* 25 (Summer 1994):6–12

Ledoit, O., "A Well-Conditioned Estimator for Large Dimensional Covariance Matrices." University of California Los Angeles. Working Paper. October 1996

Ledoit, O., "Improved Estimation of the Covariance matrix of Stock Returns with an Application to Portfolio Selection." University of California Los Angeles. Working Paper. January 14, 1999

Lee, S.L., "Correlation Shifts and Real Estate Portfolio Management." *Journal of Real Estate Portfolio Management* 9 (2003):45–57

Lee, S.L., "The Relative Importance of Property Type and Regional Factors in Real Estate Returns." *Journal of Real Estate Portfolio Management* 7 (2001):159–167

Lee, S., Byrne, P., "Diversification By Sector, Region or Function? A Mean Absolute Deviation Optimisation." *Journal of Property Valuation and Investment* 16 (1998):38–56

Lee, S., Lizieri, C., "Asset Allocation, Cross-Class Correlations and the Structure of Property Returns." Paper presented to the RICS "Cutting Edge" Research Conference, Cambridge, England, September 1999

Lee, S., Stevenson, S., "Real Estate Portfolio Construction and Estimation Risk." A paper presented at the 7th European Real Estate Society Conference, Bordeaux, France, June 14–17, 2000

Lhabitant, F.-S., *Handbook of Hedge Funds*. West Sussex, England: Wiley, 2006

Liang, Y., McIntosh, W., "Measuring the Overall and Diversification Benefits of an Investment." *Real Estate Finance* 16 (Fall 1999):55–63

Louargand, M.A., "A Survey of Pension Fund Real Estate Portfolio Risk Management Practices." *Journal of Real Estate Research* 7 (Fall 1992):361–373

MacKinlay, A.C., Pastor, L., "Asset Pricing Models: Implications for Expected Returns and Portfolio Selection." *Review of Financial Studies* 13 (Winter 2000):883–916

Malizia, E.E., Simons, R.A., "Comparing Regional Classifications for Real Estate Portfolio Diversification." *Journal of Real Estate Research* 6 (Spring 1991):53–77

Maurer, R., Reiner, F., "International Asset Allocation With Real Estate Securities in a Shortfall Risk Framework: The Viewpoint of German and U.S. Investors." *Journal of Real Estate Portfolio Management* 8 (2002):27–43

Michaud, R.O., *Efficient Asset Management: A Practical Guide to Stock Portfolio Optimization and Asset Allocation*. Boston, MA: Harvard Business School Press, 1998

Miles, M., Cole, R., Guilkey, D., "A Different Look at Commercial Real Estate Returns." *Journal of the American Real Estate and Urban Economics Association* 18 (Winter 1990):403–430

Miles, M., McCue, T., "Historic Returns and Institutional Real Estate Portfolios." *Journal of the American Real Estate and Urban Economics Association* 10 (1982):184–198

Miles, M., McCue, T., "Diversification in the Real Estate Portfolio." *Journal of Financial Research* 7 (Spring 1984):57–68

Mueller, G.R., Laposa, S.P., "Property-Type Diversification in Real Estate Portfolios: A Size and Return Perspective." *Journal of Real Estate Portfolio Management* 1 (1995):39–50

Mueller, G.R., Ziering, B.A., "Real Estate Portfolio Diversification Using Economic Diversification." *Journal of Real Estate Research* 7 (1992):375–386

Munneke, H.J., Slade, B.A., "An Empirical Study of Sample Selection Bias in Indices of Commercial Real Estate." *Journal of Real Estate Finance and Economics* 21 (July 2000):45–64

Myer, F.C.N., Webb, V.R., "Estimating allocations for mixed asset Portfolios using Bootstrap Technique." Paper presented at the American Real Estate Society Meetings, Saroscta, Florida, April 1991.

National Council of Real Estate Investment Fiduciaries. "Frequently Asked Questions About NCREIF and the NCREIF Property Index (NPI)."

Newell, G., MacFarlane, J., "Improved Risk Estimation Using Appraisal-Smoothed Real Estate Returns." *Journal of Real Estate Portfolio Management* 1 (1995):51–57

Norman, E.J., Sirmans, G.S., Benjamin, J.D., "The Historical Environment of Real Estate Returns." *Journal of Real Estate Portfolio Management* 1 (1995):1–24

Olaleye, A., Aluko, B.T., "Evaluating Managers Diversification of Real Estate Portfolios: Evidence from Nigeria." *International Journal of Strategic Property Management* (September 2007)

Ong, S.-E., Ranasinghe, M., "Portfolio Variance and Covariance Matrices." *Journal of Real Estate Portfolio Management* 6 (2000):1–6

Pagliari, J.L., Jr., (ed) *The Handbook of Real Estate Portfolio Management*. Chicago, IL: Richard D. Irwin, Inc., 1995

Pagliari, J.L., Jr., Lieblich, F., Schaner, M., Webb, J.R., "Twenty Years of the NCREIF Property Index." Working Paper, September 1998

Pagliari, J.L., Jr., Webb, J.R., Del Casino, J.J., "Applying MPT to Institutional Real Estate Portfolios: The Good, the Bad, and the Ugly." *Journal of Real Estate Portfolio Management* 1 (1995):67–88

Patel, K., Sing, T.F., "Implied Volatility in U.K. Commercial Property Market: Empirical Evidence Based on Transaction Data." *Journal of Real Estate Finance and Economics* 20 (January 2000):5–24

Robichek, A.A., Cohn, R.A., Pringle, J.J., "Returns on Alternative Investment Media and Implications for Portfolio Construction." *Journal of Business* 55 (July 1972):427–443

Roulac, S.E., "How to Structure Real Estate Investment Management." *Journal of Portfolio Management* 8 (Fall 1981):32–35

Sanders, A.B., Pagliari, J.L., Webb, J.R., "Portfolio Management Concepts and the Application to Real Estate." In: Pagliari, J.L., (ed) *Handbook of Real Estate Portfolio Management*, Chicago, IL: Irwin Professional Publishing, 1995, pp 117–172

Schuck, E.J., "Why Diversification Does Work." *Real Estate Review* 25 (Spring 1995):12–18

Schwann, G.M., "A Real Estate Price Index for Thin Markets." *Journal of Real Estate Finance and Economics* 16 (May 1998):269–287

Seiler, M.J., Webb, J.R., Myer, F.C.N., "Diversification Issues in Real Estate Investment." *Journal of Real Estate Literature* 7 (1999):163–179

Seiler, M.J., Webb, J.R., Myer, F.C.N., "Can Private Real Estate Portfolios Be Rebalanced/Diversified Using Equity REIT Shares?" *Journal of Real Estate Portfolio Management* 7 (2001):25–41

Shaoqun, W., Ying, H., "Potential for Portfolio Diversification Across China's Real Estate Markets." Paper presented at the Asian Real Estate Society Conference, Dehli, India, August 9–12, 2004

Sing, T.F., Ong, S.E., "Asset Allocation in a Downside Risk Framework." *Journal of Real Estate Portfolio Management* 6 (2000):213–223

Sirmans, C.F., Fisher, J.D., "Real Estate in Pension Fund Portfolios: A Review of the Literature and Annotated Bibliography." Pension Real Estate Association

Sivitanides, P.S., "A Downside-risk Approach to Real Estate Portfolio Structuring." Journal of Real Estate Portfolio Management 4 (1998):159–168

Stevenson, S., "Bayes-Stein Estimators and International Real Estate Allocation." Journal of Real Estate Research 21 (2001):89–103

Stevenson, S., "Ex-Ante and Ex-Post Performance of Optimal REIT Portfolios." *Journal of Real Estate Portfolio Management* 8 (2002):199–207

Thomson, T.A., "Long-Term Portfolio Returns from Timber and Financial Assets." *Journal of Real Estate Porfolio Management* 3 (1997):57–73

Thomson Financial Venture Economics. http://vx.thomsonib.com

Viezer, T.W., "Building Real Estate Portfolios One Deal at a Time, With an Eye on Diversification." *Real Estate Finance* (Fall 1999a):1–11

Viezer, T.W., "Econometric Integration of Real Estate's Space and Capital Markets." *Journal of Real Estate Research* 18 (1999b):503–519

Viezer, T.W., "Simulating and Testing Metropolitan Area/Property Type Returns." Real Estate Finance (Fall 1999c):1–16

Viezer, T.W., "Evaluating 'Within Real Estate' Diversification Strategies." *Journal of Real Estate Portfolio Management* 6 (2000):75–95

Viezer, T.W., "Statistical Strategies for Real Estate Portfolio Diversification." Doctoral Dissertation, The Ohio State University, 1998

Webb, J.R., "Real Estate Investment Acquisition Rules for Life Insurance Companies and Pension Funds: A Survey." *Journal of the American Real Estate and Urban Economics Association* 12 (Winter 1984):495–520

Webb, JR., Curico, R.J., Rubens, J.H., "Diversification Gains From Including Real Estate in Mixed-Asset Portfolios." *Decision Sciences* 19 (1988):434–452

Webb, J.R., Rubens, J.H., "Portfolio Considerations in the Valuation of Real Estate." *Journal of the American Real Estate and Urban Economics Association* 14 (1986):465–495

Webb, J.R., Rubens, J.H., "How Much in Real Estate? A Surprise Answer." *Journal of Portfolio Management* 13 (1987):10–14

Wellner, K., Thomas, M., "Diversification Benefits From European Direct Real Estate Investments With a Special Focus on the German Market." Paper presented at the Annual Meeting of the American Real Estate Society, Captiva Iceland, 2004

Wiley, R.J., "Real Estate Investment Analysis: An Empirical Study." *Appraisal Journal* (October 1976):586–592

Wilson, P.J., Zurbruegg, R., "International Diversification of Real Estate Assets: Is It Worth It? Evidence from the Literature." *Journal of Real Estate Literature* 11 (2003):259–277

Worzala, E.M., Bajelsmit, V.L., "Real Estate Asset Allocation and the Decisionmaking Framework Used By Pension Fund Managers." *Journal of Real Estate Portfolio Management* 3 (1997):46–56

Young, M.S., "Making Sense of the CNREIF Property Index: A New Formulation Revisited." *Journal of Real Estate Portfolio Management* 11 (2005):211–223

Young, M.S., "Revisiting Non-Normal Real Estate Return Distributions by Property Type in the U.S." *Journal of Real Estate Finance and Economics*. 36 (February 2008):233–248

Young, M.S., Graff, R.A., "Real Estate is Not Normal: A Fresh Look at Real Estate Return Distributions." *Journal of Real Estate Finance and Economics* 10 (May 1995):225–259

Young, M.S., Greig, D.W., "Drums Along the Efficient Frontier." *Real Estate Review* 22 (Winter 1993):18–29

Young, M.S., Lee, S.L., Devaney, S.P., "Non-Normal Real Estate Return Distributions by Property Type in the U.K." *Journal of Property Research* 23 (June 2006):109–133

Zerbst, R.H., Cambon, B.R., "Real Estate: Historical Returns and Risks." *Journal of Portfolio Management* 10 (Spring 1984):5–20

Ziering, B., Hess, R., "A Further Note on Economic Versus Geographic Diversification." *Real Estate Finance* (Fall 1995):53–60

Ziobrowski, B.J., Ziobrowski, A.J., "Higher Real Estate Risk and Mixed-Asset Portfolio Performance." *Journal of Real Estate Portfolio Management* 3 (1997):107–115

ERRATUM TO:

Foreword

Harry Markowitz

J.B. Guerard (ed.), *Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*,
DOI 10.1007/978-0-387-77439-8, © Springer Science+Business Media, LLC 2010

DOI 10.1007/978-0-387-77439-8_30

The foreword was published without the following critical paragraph:

Sharpe (1964) and Merton (1990) versus Jacobs, Levy and Markowitz (2004, 2010)

Sharpe (1964) and Lintner (1965) present an "equilibrium" model. They say that, given certain assumptions, "in equilibrium" such-and-such will be true. Their model may be interpreted as a single-period or a static steady-state model. On the other hand, Merton (1990) and his many followers present continuous-time models in which price is assumed to follow one or another stochastic process, assumed *a priori*. In contrast to both of these types of models—the static and the continuous-time dynamic—the model presented by Jacobs, Levy and Markowitz (2004, 2010) is an asynchronous discrete event simulation in which time advances, usually in irregular jumps, to the next most imminent event. Prices are endogenous, resulting from the interaction of thousands of investors and their traders following various investment and trading rules.

Harry Markowitz

References

Jacobs, B.I., K.N. Levy and H.M. Markowitz (2004) "Financial Market Simulation." *The Journal of Portfolio Management*. 30th Anniversary Issue:142-51.

Jacobs, B.I, K.N. Levy and H.M. Markowitz (2010) "Simulating Security Markets in Dynamic and Equilibrium Modes." *Financial Analysts Journal*.

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* 47:13-37.

Merton, R.C. (1990) *Continuous-Time Finance*. Blackwell Publishing Ltd., Cambridge.

Sharpe, W.F. (1964) "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk." *The Journal of Finance* 9(3):425-42.

The online version of the original chapter can be found under
DOI 10.1007/978-0-387-77439-8

About the Editor and Contributors

About the Editor

John B. Guerard, Jr., Ph.D., is the Director of Quantitative Research at McKinley Capital Management, in Anchorage, AK. He earned his AB in economics from Duke University, MA in economics from the University of Virginia, MSIM from the Georgia Institute of Technology, and Ph.D. in Finance from the University of Texas, Austin. John taught at the McIntire School of Commerce, the University of Virginia, Lehigh University, and Rutgers University. John taught as an adjunct faculty member at the International University of Monaco and the University of Pennsylvania. He worked with the DAIS Group at Drexel, Burnham, Lambert, Daiwa Securities Trust Company, Vantage Global Advisors, and served on the Virtual Research team at GlobeFlex Capital. John co-managed a Japanese equity portfolio with Harry Markowitz at Daiwa Securities Trust Company. While serving as the Director of Quantitative Research at Vantage Global Advisors (formerly MPT Associates), Mr. Guerard was awarded the first Moskowitz Prize for research in socially responsible investing. Mr. Guerard has published several monographs, including *The Handbook of Financial Modeling* (Probus, 1989, with H.T. Vaught), *Corporate Financial Policy and R&D Management* (Wiley, 2006, second edition), and *Quantitative Corporate Finance* (Springer, 2007, with Eli Schwartz). John serves as an Associate Editor of the *Journal of Investing* and *The International Journal of Forecasting*. Mr. Guerard has published research in *The International Journal of Forecasting*, *Management Science*, the *Journal of Forecasting*, *Journal of Investing*, *Research in Finance*, *Research Policy*, and the *Journal of the Operational Research Society*.

About the Contributors

Christopher R. Blake holds the Joseph Keating, S.J., Distinguished Professor chair at Fordham University's Graduate School of Business Administration and College of Business Administration in New York City. He joined the Fordham faculty in 1991. Dr. Blake has published articles in *The European Finance Review*, *The Journal of Business*, *The Journal of Finance*, *The Journal of Financial and Quantitative Analysis*, *The Journal of Financial Intermediation*, *The Journal of Public Economics*, *The Review of Economics and Statistics*, and *The Review of Financial Studies*. He is a co-developer of *The Investment Portfolio*, Windows-based portfolio-management software, and he has consulted in the areas of investments and portfolio management with major financial and governmental institutions.

Efim M. Bronshtein is a Professor at Ufa State Aviation Technical University (Russia) in Mathematics and Financial Mathematics. He holds a Ph.D. (1980) from the Institute of Mathematics in Siberian Branch of Soviet Academy of Sciences and a Sc.D. (1999) from Bashkirian State University. He is a member of the European Working Group "Financial Modelling" (EWG FinMod) and a Scientist in Honour of Republik of Bashkortostan (2007).

Andrew H. Chen is a Distinguished Professor of Finance at the Cox School of Business, SMU. He received his B.A. in economics from National Taiwan University, M.A. in economics and Ph.D. in finance from UC, Berkeley. He has edited and co-authored many books, and his numerous articles have appeared in top academic and professional journals including *Journal of Finance*, *Journal of Financial Economics*, *Journal of Business*, *Journal of Financial and Quantitative Analysis*, *Management Science*, *Financial Management*, *Journal of Portfolio Management*, *European Journal of Political Economy*, *Journal of Money*, *Credit, and Banking*, and others. He has served as the Editor of *Research in Finance*, and a Managing Editor of the *International Journal of Theoretical and Applied Finance.* He has served as chairman, director, and advisory director of several financial institutions and corporations. He has been a visiting scholar at prominent universities in Hong Kong, Singapore, Taiwan, and Australia. He is a former President of the Financial Management Association International.

Meenakshi Sundaram Chettiappan is a Quantitative Research Analyst at McKinley Capital Management, LLC, Anchorage. He earned his M.S. Quantitative Computational Finance at the Georgia Institute of Technology and a Bachelor of Engineering, Computer Science and Engineering, Anna University.

Andrew Clark was recently appointed as the Chief Index Strategist and the Head of Research for Thomson Reuters Indices, a new business venture of Thomson Reuters. In this new role, Andrew has the responsibility of developing new index methodologies, vetting index methodologies proposed by partners of TRI, making sure the indices perform as stated in the documentation after they are released, and being the chief "evangelist" for the Thomson Reuters Index business. Prior to his current role, Andrew was the Head of Research for the Americas for Lipper and Global Manager of Quantitative Analysis for Reuters' Research and Asset Management division. In this role, he was also the main spokesperson for Lipper in Canada, the USA, and Latin America. Andrew has made many appearances on CNBC, BNN, MSNBC in Hong Kong, CCTV, and CNNfN. He has spoken at many industry, economic, and mathematical conferences where he has presented original research and discussed trends in the market and the outlook for markets world-wide. Andrew has won several awards for his research – most recently the Best Investment Paper award from the Midwest Finance Association – and is regularly quoted in the *Wall Street Journal*, the *New York Times*, the *International Herald Tribune*, and many magazines and papers around the world. His research work has appeared in *Physica A*, *Quantitative Finance*, *European Physical Journal B*, and many other research journals. In his quantitative research role for Research and Asset Management, Andrew helped lead the exploration of new businesses for Reuters in such areas as the development of investible actively managed indices, and the building of better credit rating tools for asset back securities and corporate bonds. He also worked with the Chinese Ministry of Labor and the Chinese government's Social Security Fund to develop "best practices" guidelines and benchmarks for investment companies in China.

Gregory Connor is the Professor of Finance at the National University of Ireland, Maynooth. He previously taught at the London School of Economics (1991–2008), University of California, Berkeley (1986–1991) and Northwestern University (1981–1986). He obtained his Ph.D. (economics) from Yale University and his B.A. (economics) from Georgetown University. He has published widely on multi-factor pricing models, financial econometrics, and security market pricing models.

Richard W. Cottle is an emeritus professor of Management Science and Engineering in the School of Engineering at Stanford University. A graduate of Harvard University (A.B. 1957 and A.M. 1958), Cottle received his Ph.D. from the University of California, Berkeley in 1964, all three degrees being in mathematics. After 3 years as a Member of the Technical Staff at Bell Laboratories (Holmdel, NJ), Cottle joined the faculty of the Operations Research Department at Stanford in 1967 as an assistant professor. He became a full professor in 1973 and served as department chair from 1990 to 1996. Cottle spent a year (1978–1979) at the Universities of Bonn and Cologne as recipient of a U.S. Senior Scientist award from the Alexander von Humboldt Foundation in Bonn, Germany. For their monograph The Linear Complementarity Problem, Cottle (and his co-authors) shared the 1994 Lanchester Prize from INFORMS for the best work published in English in Operations Research. In 2002, he spent a term at the Judge Institute of Management Studies, University of Cambridge, as the Arthur Anderson Distinguished Visitor. He became an INFORMS Fellow in 2006. Cottle's research interests lie within the area of mathematical programming, primarily linear, quadratic and nonlinear optimization and linear complementarity theory. His research publications appear in *Mathematical Programming*, *Linear Algebra and its Applications*, *SIAM Journal on Applied Mathematics*, the *Journal on Global Optimization*, and several others. In addition to the aforementioned monograph (to be re-issued in SIAM's Classics Series), Cottle's publication list includes *Stanford Street Names: A Pocket Guide* and ten edited (or co-edited) books including *The Basic George B. Dantzig*. Other works in progress include a textbook, *Linear and Nonlinear Programming*, a text/reference, *Quadratic Programming*, papers on the linear complementarity problem, and a history of the International Symposia on Mathematical Programming.

Ran Duchin is an Assistant Professor of Finance at the Ross School of Business, University of Michigan, Ann Arbor. He received a B.A. from Hebrew University (2001) and Ph.D. from the University of Southern California (2008). His research areas of interest include portfolio selection, corporate finance, and corporate governance.

Edwin J. Elton is a Nomura Professor of Finance at the Leonard N. Stern School of Business, New York University. Professor Elton has authored or coauthored eight books and over 115 articles. The seventh edition of his book, *Modern Portfolio Theory and Investment Analysis*, is the standard textbook used in most leading graduate schools of business. His collected works in investments are published in two volumes by MIT Press. He was formerly co-managing editor of the *Journal of Finance*. He has been a member of the Board of Directors of the American Finance Association and an associate editor of *Management Science*. Professor Elton has served as a portfolio theory and investment management consultant for many major financial institutions in Asia, Europe, and the United States. He has been a senior research fellow at the International Institute of Management in Berlin, visiting scholar and is on the faculty at the European Institute for Advanced Studies in Management (EIASM) in Brussels and visiting professor at Katholieke Universiteit Leuven. Professor Elton received the Graham Dodd award for research in investments and was named Distinguished Scholar by the Eastern Finance Association. He received the James Vertin lifetime achievement award from financial analyst federation. He is a former president of the American Finance Association (1996) and a fellow of the society.

Frank J. Fabozzi is a Professor in the Practice of Finance and Becton Fellow at the Yale School of Management. Prior to joining the Yale faculty, he was a Visiting Professor of Finance in the Sloan School at MIT. Professor Fabozzi is a Fellow of the International Center for Finance at Yale University and on the Advisory Council for the Department of Operations Research and Financial Engineering at Princeton University. He is an Affiliated Professor at the University of Karlsruhe (Germany) Institute of Statistics, Econometrics and Mathematical Finance. He earned a doctorate in economics from the City University of New York in 1972. In 2002, he was inducted into the Fixed Income Analysts Society's Hall of Fame and is the 2007 recipient of the C. Stewart Sheppard Award given by the CFA Institute. He is the editor of the *Journal of Portfolio Management*.

Christopher G. Green is a doctoral candidate in Statistics at the University of Washington. He specializes in risk analysis and factor models. He holds a Masters in Mathematics from the University of Washington.

Martin J. Gruber, Ph.D. , is a Nomura Professor of Finance at the Leonard N. Stern School of Business, New York University. He is a co-author of the leading graduate level text *Modern Portfolio Theory and Investment Analysis*, currently in its seventh edition. He has published more than 100 articles on investment management in such journals as the *Journal of Finance*,*Management Science*,*Journal of Business*,*Journal of Financial and Quantitative Analysis*,*Operations Research*,*Oxford Economic Papers*, and the *Financial Analyst Journal*. Professor Gruber is the past president of the American Finance Association and a fellow of both the American Finance Association and the Financial Management Association. He has been a co-editor of the *Journal of Finance* and department editor of Finance for *Management Science*. Professor Gruber is a member of the Board and the executive committee of the National Bureau of Economic Research. He is also a member of the board of the Daiwa Closed End Mutual Funds. He was previously a member of the board of TIAA-CREF, chairman of the board of CREF, a member of the board and chairman of the Equity Committee of DWS Mutual Funds and a member of the board of SGCowen Mutual Funds.

Robert A. Haugen has held endowed professorships at the Universities of Wisconsin, Illinois, and California. He has written 15 books, including *The Incredible January Effect* and *The New Finance*, which have collectively been published in eight languages. He has been featured in *Barrens, Money, Smart Money*, and has appeared in most major media.

Dashan Huang is a doctoral student in Finance at Olin Business School, Washington University in St. Louis. His current research interests include empirical asset pricing and portfolio allocation. He received a doctorate in Engineering from Kyoto University in 2007. His research has appeared in *Journal of Economic Dynamics and Control*, *Annals of Operations Research*, *Energy Economics*, and *European Journal of Operational Research*.

Gerd Infanger is a Consulting Professor in the Department of Management Science and Engineering (formerly Operations Research) at Stanford University and Founder and CEO of Infanger Investment Technology, LLC, a federally registered investment advisory company in Mountain View, California. Infanger received his Master's degree in electrical engineering and economics at Graz University of Technology in 1983 and his Ph.D. with honors at Vienna University of Technology in energy economics and operations research in 1986. He received tenure as an Associate Professor (Univ.-Doz.) at Vienna University of Technology in 1993. At the beginning of 1989, Infanger came to visit the Department of Operations Research at Stanford University, collaborating with George B. Dantzig in the field of optimization under uncertainty. Since then, Infanger's research has concentrated on the development of theory, algorithms and software for solving large-scale stochastic programs. In 1998, based on his proprietary quantitative methods of equity performance prediction, Infanger founded Infanger Investment Technology, LLC, with the goal to provide state-of-the-art quantitative investment advisory services and software offerings to institutional investors. At the beginning of 2008, the firm advised over 300 million dollars in assets in various global tactical asset allocation funds and American and European equity funds. Infanger has published a book: *Planning under Uncertainty: Solving Large-Scale Stochastic Programs* as well as articles in various journals, including *Annals of Operations Research*, *Mathematical Programming*, *Quantitative Finance*, and various book chapters including a recent chapter about dynamic portfolio strategies in the *Handbook of Asset and Liability Management*. With respect to software, Infanger is the principal author of DECIS, a system for solving large-scale stochastic programs, IITPortf™, a large-scale portfolio optimization system, and WealthiOR™, a web-based software system for dynamic asset allocation. Infanger's research focuses on stochastic and dynamic programming, particularly as applied to portfolio optimization and asset allocation, as well as other quantitative aspects of asset management.

Bruce I. Jacobs is co-founder and principal of Jacobs Levy Equity Management, a leading provider of quantitative equity strategies. He is the co-chief investment officer, portfolio manager, and co-director of research. Dr. Jacobs's articles on equity management have appeared in the *Financial Analysts Journal*, *Journal of Portfolio Management*, *Journal of Investing*, *Japanese Security Analysts Journal*, and *Operations Research*. He has received *Financial Analysts Journal* Graham and Dodd Awards and outstanding article awards from the *Journal of Portfolio Management* and *Journal of Investing*. Dr. Jacobs is the author of *Capital Ideas and Market Realities: Option Replication, Investor Behavior, and Stock Market Crashes* (Blackwell), co-author with Ken Levy of *Equity Management: Quantitative Analysis for Stock Selection* (McGraw-Hill and a Chinese translation), co-editor with Ken Levy of *Market Neutral Strategies* (Wiley), and co-editor of *The Bernstein Fabozzi/Jacobs Levy Awards: Five Years of Award-Winning Articles from the Journal of Portfolio Management* (Institutional Investor). He was a featured contributor to *How I Became a Quant: Insights from 25 of Wall Street's Elite* (Wiley). Dr. Jacobs has spoken at many forums, including the Institute for Quantitative Research in Finance, Berkeley Program in Finance, CFA Institute, Rutgers University, Society of Quantitative Analysts, and New York Society of Security Analysts, and he has given a *Financial Analysts Journal* Media Seminar and presented at conferences for Goldman Sachs and Morgan Stanley.

Formerly he was the first Vice President of the Prudential Insurance Company of America, where he served as Senior Managing Director of a quantitative equity management affiliate of the Prudential Asset Management Company and Managing Director of the discretionary asset allocation unit. Prior to that, he was on the finance faculty of the University of Pennsylvania's Wharton School and consulted to the Rand Corporation. Dr. Jacobs has a B.A. from Columbia College, an M.S. in Operations Research and Computer Science from Columbia University's School of Engineering and Applied Science, an M.S.I.A. from Carnegie Mellon University's Graduate School of Industrial Administration, and an M.A. in Applied Economics and a Ph.D. in Finance from the Wharton School. He is an Associate Editor of the *Journal of Trading* and serves on the *Journal of Portfolio Management* Advisory Board and the *Financial Analysts Journal* Advisory Council.

Ronald N. Kahn is the Managing Director of Barclays Global Investors. Ron Kahn heads BGI's efforts to develop, innovate, and sustain BGI's active equity strategies. His roles at BGI have included a prior stint as Global Head of Equity Research, plus Global Head of Advanced Equity Strategies, and Head of Active Equities in the USA. Prior to joining BGI, Ron spent more than 7 years as the director of research at BARRA, where his research spanned the equity and fixed income markets in the USA and globally. He co-authored, with BGI's Richard Grinold, the influential book

Active Portfolio Management: Quantitative Theory and Applications. He is a 2007 winner (along with BGI's Matt Scanlan, and Larry Siegel of the Ford Foundation) of the Bernstein Fabozzi/Jacobs Levy award for the best article in the *Journal of Portfolio Management*. He serves on the editorial advisory boards of the *Journal of Portfolio Management* and the *Journal of Investment Consulting*. The 2007 book, *How I Became a Quant*, includes his essay describing his transition from physics to finance. Ron received a Ph.D. in physics from Harvard University in 1985 and an AB in physics, summa cum laude, from Princeton in 1978, and was a post-doctoral fellow in physics at UC Berkeley.

Woo Chang Kim is a Ph.D. candidate in the Operations Research and Financial Engineering Department at Princeton University. He holds a bachelor's degree in industrial engineering from the Seoul National University and has taught at the Korean Naval Academy as a lecturer. Woo Chang's main areas of interest are financial planning and risk management. His work has been focused on practical issues for long-term investors such as pension plans as well as computational issues in large-scale stochastic optimization problems. He has published in the *Journal of Portfolio Management* and the *Journal of Financial Transformation*.

Cheng-few Lee is a Distinguished Professor of Finance at Rutgers Business School, Rutgers University and was the chairperson of the Department of Finance from 1988 to 1995. He has also served on the faculty of the University of Illinois (IBE Professor of Finance) and the University of Georgia. He has maintained academic and consulting ties in Taiwan, Hong Kong, China, and the United States for the past three decades. He has been a consultant to many prominent groups including, the American Insurance Group, the World Bank, the United Nations, The Marmon Group Inc., Wintek Corporation, Polaris Financial Group, and others. Professor Lee founded the *Review of Quantitative Finance and Accounting* (RQFA) in 1990 and the *Review of Pacific Basin Financial Markets and Policies* (RPBFMP) in 1998, and serves as a managing editor for both journals. He was also a co-editor of the *Financial Review* (1985–1991) and the *Quarterly Review of Economics and Business* (1987–1989). In the past 36 years, Dr. Lee has written numerous textbooks ranging in subject matters from financial management to corporate finance, security analysis and portfolio management to financial analysis, planning and forecasting, and business statistics. In addition, he has edited a popular book entitled *Encyclopedia of Finance* (with Alice C. Lee). Dr. Lee has also published more than 170 articles in more than 20 different journals in finance, accounting, economics, statistics, and management. Professor Lee has been ranked the most published finance professor worldwide during 1953–2002. Professor Lee was the intellectual force behind the creation of the new Masters of Quantitative Finance program at Rutgers University. This program began in 2001 and has been ranked as one of the top 10 quantitative finance programs in the United States.

Haim Levy is the Dean of School of Business in the center of Law and Business, Israel. He holds a B.A. from Hebrew University (1963), M.A., Hebrew University (1966), and Ph.D., Hebrew University (1969), and has received the Hebrew

University Presidential Distinguished Researcher Award (1966) and the EMET Prize (2006). He has published numerous papers in the *Journal of Finance*, *Journal of Financial and Quantitative Analysis*, *Review of Financial Studies*, *Management Science*, *American Economic Review*, *Review of Economic Studies*, *Econometrica*, and *Journal of Economic Theory.* Ranked as the most prolific researcher in Finance (in 16 leading journals in the area of Finance), for the period 1945–1986, and for the period 1953–2002, his research areas of interest include portfolio selection, stochastic dominance, international finance, and behavioral economics.

Kenneth N. Levy is a co-founder and principal of Jacobs Levy Equity Management, a leading provider of quantitative equity strategies. He is the co-chief investment officer, portfolio manager, and co-director of research. His articles on equity management have appeared in the *Financial Analysts Journal*, *Journal of Portfolio Management*, *Journal of Investing*, *Japanese Security Analysts Journal*, and *Operations Research*. He has received *Financial Analysts Journal* Graham and Dodd Awards and the *Journal of Portfolio Management* outstanding article award. Ken Levy is a co-author with Bruce Jacobs of *Equity Management: Quantitative Analysis for Stock Selection* (McGraw-Hill and a Chinese translation), co-editor with Bruce Jacobs of *Market Neutral Strategies* (Wiley), and co-editor of *The Bernstein Fabozzi/Jacobs Levy Awards: Five Years of Award-Winning Articles from the Journal of Portfolio Management* (Institutional Investor). He was a featured contributor to *How I Became a Quant: Insights from 25 of Wall Street's Elite* (Wiley). Ken has lectured at the Wharton School and spoken at various forums, including the Institute for Quantitative Research in Finance, Berkeley Program in Finance, CFA Institute, Rutgers University, Society of Quantitative Analysts, and Corporate Earnings Analysis Seminar. Formerly he was Managing Director of a quantitative equity management affiliate of the Prudential Asset Management Company. Prior to that, he was responsible for quantitative research at Prudential Equity Management Associates. Ken has a B.A. in Economics from Cornell University, an M.B.A. and an M.A. in Business Economics from the University of Pennsylvania's Wharton School, and completed all requirements short of the dissertation for a Ph.D. at Wharton. He is a CFA charterholder and has served on the CFA Candidate Curriculum Committee, POSIT Advisory Board, and the investment board of a community foundation.

R. Douglas Martin is a Professor of Statistics, Adjunct Professor of Finance, and Director of Computational Finance at the University of Washington, and a former Chairman of the Department of Statistics. He was a consultant in the Mathematics and Statistics Research Center at Bell Laboratories for 10 years, a position created on the recommendation of J.W. Tukey, whose influence resulted in Martin switching his research focus from Electrical Engineering to Statistics. In 1987, Martin founded Statistical Sciences, Inc. to develop and market the S-PLUS system for data analysis and statistical modeling, based on an exclusive publishing license for the S language from Bell Laboratories. In 1993, he sold StatSci to MathSoft, Inc. and remained with MathSoft, subsequently renamed Insightful, Inc., in various roles until 2003. Subsequently, Martin was a founder and Chairman of FinAnalytica, Inc., developer of the Cognity portfolio construction and risk management system, where he served as CEO from 2006 to 2008. He has authored numerous publications in the areas of time series analysis and robust statistical methods, including two Royal Statistical Society Discussion papers and one invited Annals of Statistics Discussion paper. His recent research focus has been on applications of robust statistical methods, fat-tailed skewed distributions, and downside risk measures in portfolio construction and risk management. Martin is a co-author of *Modern Portfolio Optimization* (2005, with Bernd Scherer, Springer) and *Robust Statistics: Theory and Methods* (2006, with Ricardo Maronna and Victor Yohai, Wiley). He holds the B.S.E. and Ph.D. degrees in Electrical Engineering from Princeton University.

Jose Menchero is the Executive Director and Senior Researcher at MSCI Barra, where he focuses on global equity risk modeling and portfolio analytics. Since joining MSCI Barra in January of 2007, Jose developed a general framework for attributing return, risk, and information ratio to a set of custom investment factors. He also led the research team that developed the Barra Global Equity Model, GEM2. Before joining MSCI Barra, Jose was the Head of Quantitative Research at Thomson Financial, where he worked on performance attribution, risk attribution, and factor risk modeling. Jose has several publications in these areas. Prior to entering finance, Jose was a Professor of Physics at the University of Rio de Janeiro, Brazil. His area of research was in the Quantum Theory of Solids, and he also has several publications in this field. Jose serves on the Advisory Board of the Journal of Performance Measurement. He holds a B.Sc. degree in Aerospace Engineering from the University of Colorado at Boulder, and a Ph.D. degree in Theoretical Physics from the University of California at Berkeley. Jose is also a CFA charterholder.

Andrei Morozov is a Vice President at MSCI Barra. Andrei joined the firm in October 2005, and focuses on global equity risk modeling. Prior to joining MSCI Barra, Andrei pursued diverse research interests ranging from experiments in electron coincidence spectroscopy at the Max Plank Institute of Microstructure Physics in Germany to leading research and development of proprietary video compression algorithms at XVD Corp, a technology startup in Silicon Valley. Andrei earned a Ph.D. degree in Physics from Saint Petersburg State University in Russia.

John M. Mulvey is a Professor of Operations Research and Financial Engineering and a founding member of the Bendheim Center for Finance at Princeton University. He has been a faculty member at Princeton for almost 31 years. He has supervised over 20 doctoral students, and close to 150 senior theses at Princeton. His specialty is large-scale optimization, with applications to strategic financial planning and other domains. He has implemented asset-liability management systems for many companies, including PIMCO, American Express, Towers Perrin and Tillinghast, Merrill Lynch, American Re-Insurance, and others. These systems integrate key risks and opportunities to assist the organization in making critical high-level decisions. His current research involves the role of alternative investments for institutional investors, such as pension plans, insurance companies, university endowments, and family trusts, with an emphasis on approaches for optimizing performance and protecting the investor's wealth over time. He has invented a systematic approach for investing, called Dynamic Portfolio Theory. This theory emphasizes flexibility and responsiveness, especially during periods of high turbulence.

Andras Niedermayer is a postdoctoral research fellow at Northwestern University. He graduated with a licenciatus in Economics and Computer Sciences at the University of Bern in 2002 and with a Master of Science in Economics at the London School of Economics in 2003. In 2008 he received his Ph.D. in economics (summa cum laude) at the University of Bern. He is the recipient of a Swiss National Science Young Researcher Grant and a NET Institute Research Grant. Andras Niedermayer's research interests are in computational economics and finance, market microstructure, mechanism design, and applied microeconomic theory.

Daniel Niedermayer graduated in economics and computer sciences at the University of Bern. He received his Ph.D. in finance (summa cum laude) from the University of Basel in 2008. During his studies, he was working as a software developer and as a part-time teaching assistant for his thesis committee chair Prof. Heinz Zimmermann. Currently, Daniel Niedermayer works for Quantitative Portfolio Engineering within Credit Suisse Asset Management (MACS) in Zurich. His research interests include robust portfolio optimization and applications in computational finance. He was given the Schmeller-Fonds award of the Economic Society of the Canton Bern.

Svetlozar T. Rachev is a co-founder and the President of BRAVO Risk Management Group - originator of the Cognity methodology, which was acquired by FinAnalytica. He holds the Chair-Professorship in Statistics, Econometrics, and Mathematical Finance at University of Karlsruhe, and is the author of 12 books and over 300 published articles on finance, econometrics, statistics, and actuarial science. At University of California at Santa Barbara, he founded the Ph.D. program in mathematical and empirical finance. Professor Rachev holds Ph.D. (1979) and Doctor of Science (1986) degrees from Moscow University and Russian Academy of Sciences, under the supervision of A. Kolmogorov, Yu. Prohorov and L. Kantorovich. His scientific work lies at the core of Cognity's newer and more accurate methodologies in risk management and portfolio analysis.

Boryana Racheva-Iotova is the President of FinAnalytica. She leads FinAnalytica's R&D team and product development of the Cognity platform which offers risk and portfolio construction analytics for the Fund of Funds and Hedge Fund industries. She was a co-founder and CEO of the Bravo Risk Management Group – the originator of Cognity product suite, which was later acquired by FinAnalytica. She holds a M.Sci. in Probability and Statistics at the Faculty of Mathematics and Informatics, Sofia University. Her many articles have recently appeared in premier scientific journals. She has nearly 10 years of experience in building risk management solutions and utilizing latest analytical advancements to meet the needs of the financial industry practitioners.

Martin Savelsbergh is Schneider professor in the H. Milton Stewart School of Industrial and Systems Engineering of Georgia Institute of Technology. He is an optimization specialist with over 20 years of experience in mathematical modeling, operations research, optimization methods, algorithm design, performance analysis, supply chain management, production planning, and transportation. He has published over 100 research papers in many of the top operations research and optimization journals. Dr. Savelsbergh holds a Ph.D. from Erasmus University Rotterdam and a M.Sc. from the University of Amsterdam. Dr. Savelsbergh is a founding partner of Axioma, Inc., a leading provider of tools for implementing investment strategies. At Axioma, he focuses on product innovation driven by optimization.

Ekaterina N. Sereda is a student in the Masters Program in Financial Engineering at the HECTOR School of Engineering and Management, University of Karlsruhe (Germany). She received a diploma from Ufa State Aviation Technical University (Russia) in 2007.

Peter Shepard is the Vice President and Senior Researcher at MSCI Barra, where he focuses on global equity and multi-asset class risk modeling. He was an architect of the Barra Global Equity Model, GEM2, and is the lead researcher for the Barra Integrated Model. Prior to joining MSCI Barra, he worked in fixed income factor modeling in the quantitative research group at Thomson Financial. Dr. Shepard holds a Ph.D. in theoretical physics from the University of California at Berkeley, where he researched string theory and the quantum theory of gravity. He has publications in theoretical physics and finance. Dr. Shepard also holds a Bachelors degree in physics and mathematics from Brown University.

Bernell K. Stone is the Harold F. Silver Professor of Finance, The Marriott School of Management, Brigham Young University. He has degrees from Duke University (B.S., physics), University of Wisconsin (M.S., physics, 1966), and M.I.T. (Ph.D., Applied Mathematics and Management, 1969). Before BYU, he taught at Cornell and Georgia Tech. He was the founding editor for the *Journal of Cash Management* and the founding executive director for the National Corporate Cash Management Association (now The Association for Finance Professionals). He has been an editor, co-editor, or editorial board member for eight journals including *JFQA* and *Financial Management.* He is the author of five books and more than 100 articles.

Stoyan V. Stoyanov is the Head of Quantitative Research at FinAnalytica specializing in financial risk management software. He completed his Ph.D. degree with honors in 2005 from the School of Economics and Business Engineering (Chair of Statistics, Econometrics and Mathematical Finance) at the University of Karlsruhe and is author and co-author of numerous papers. His research interests include probability theory, heavy-tailed modeling in the field of finance, and optimal portfolio theory. His articles have recently appeared in the *Journal of Banking and Finance*, *Applied Mathematical Finance*, *Applied Financial Economics*, *International Journal of Theoretical and Applied Finance*, and *Economics Letters*. He is a co-author of the mathematical finance book *Advanced Stochastic Models*, *Risk Assessment and*

Portfolio Optimization: The Ideal Risk, Uncertainty and Performance Measures. Dr Stoyanov has years of experience in applying optimal portfolio theory and market risk estimation methods when solving practical problems of clients of FinAnalytica.

Olga Streltchenko received her diploma in mechanical engineering from Kiev University in 1992, her master's in mathematics from Syracuse University in 1997, and her Ph.D. in computer science from University of Maryland Baltimore County in 2003. She is currently the Vice President at Bank of Montreal, working on a comprehensive and integrated risk reporting initiative. Her research interests lie in the area of decision support in capital markets, and include hedging illiquid liabilities with trading products, price formation of derivative securities, and multi-agent simulation.

Robert Stubbs is the Vice President, Research at Axioma, Inc., a leading provider of tools for implementing investment strategies. He joined Axioma in 2000 and is responsible for leading the research team in the design of optimization modeling and algorithmic development for portfolio construction, the development of factor risk models for equities, and the development of other analytical tools that are incorporated into Axioma's products and services. Dr. Stubbs earned a Ph.D. in Industrial Engineering and Management Science and a Master of Science degree in Industrial Engineering and Management Science from Northwestern University, and a Bachelor of Science degree in Applied Mathematics from Auburn University. Dr. Stubbs has published articles in *Mathematical Programming*, *SIAM Journal on Optimization*, *Journal of Global Optimization*, *Journal of Asset Management*, and *Institutional Investor*. Before joining Axioma, he was an Optimization Software Developer for LINDO Systems, Inc., where he worked on developing and maintaining the algorithmic components of LINDO products. Prior to that, Stubbs worked for First National Bank of Chicago and AT&T. At First National Bank of Chicago, he developed a simulation model for computing the expected distribution of the loss in the corporate credit portfolio. The efficiency of this model enabled simulation of the entire credit portfolio. At AT&T, he developed an automated scheduling system that dramatically reduced the number of machine changes needed to color fiber optic cable.

Wei Sun is Assistant Professor in the Institute of Statistics, Econometrics and Mathematical Finance at the University of Karlsruhe, Germany. He earned a doctorate in financial econometrics from the University of Karlsruhe, Germany with honors.

Makoto Suzuki is a Senior Researcher at the Daiwa Institute of Research. Mr. Suzuki has experienced in the Investment, Corporate Finance, Corporate Governance, and Pension Management. He joined Daiwa Institute in 1988 and concentrated in the research field for several years. Mr. Suzuki had worked with Harry Markowitz and John Guerard in the USA between 1992 and 1994. After he returned to Japan, he served at Pension Management Department and Investment Strategy Department for 6 years. Mr. Suzuki had organized the plan sponsors, asset management companies and consultants for making "a Risk Standard of 20 for

Japanese Pensions." He rotated to New York again in 2000. When he was working in New York, he published a book titled *A Corporate Governance in the US Pension Funds*. He covered not only the conventional investment but also alternatives in his research subjects. He was promoted as a President and CEO at an affiliate new company of Daiwa Fund Consulting in the USA. Mr. Suzuki organized and managed "The Hedge Fund Conference" with the Stern School of New York University and Daiwa Securities Group in Tokyo, 2007. Currently, Mr. Suzuki has returned to Japan and served as a senior researcher for the capital market. His main interests are Corporate Finance, Corporate Governance, Investment and Pension Management. Mr. Suzuki received B.S. from the Tohoku University (Japan), M.A. from Waseda University (Tokyo), and MBA from the Johnson Graduate School of Management, Cornell University. Mr. Suzuki contributes a lot of articles related to finance and financial business. His primary publications are *A Corporate Governance in the US Pension Funds* (February 2004, Toyokeizai) and "PSR – An efficient stock selection tool?" *Journal of International Forecasting* (June 1998, Elsevier).

Dieter Vandenbussche is a Senior Research Associate at Axioma Inc., a leading provider of tools for implementing investment strategies. His main research interests focus on computational and theoretical aspects of mathematical optimization, in particular as it applies to quantitative finance. Dr. Vandenbussche holds a Ph.D. and B.S. in Industrial Engineering from the Georgia Institute of Technology. He has published in a variety of leading journals, including *Mathematical Programming*, *SIAM Journal on Optimization*, *Journal of Global Optimization*, *INFORMS Journal of Computing*, and *Computational Optimization and Applications*. Dr. Vandenbussche was a finalist in the INFORMS Nicholson student paper competition in 2003 and was selected to attend the INFORMS doctoral colloquium in 2000. Dr. Vandenbussche has been invited to present his work at various universities and conferences, including the University of Wisconsin, Northwestern, the University of Iowa, INFORMS national meetings, and SIAM optimization conferences. Prior to joining Axioma, Dr. Vandenbussche was an assistant professor at the University of Illinois at Urbana-Champaign, where he taught graduate and undergraduate courses in Industrial Engineering and directed graduate research.

James H. Vander Weide is Research Professor at Duke University, the Fuqua School of Business. He holds a Ph.D. in finance from Northwestern University and a B.A. in Economics from Cornell University. He has published research on topics such as the optimal geometric mean portfolio strategy, optimal investment policies of regulated companies, the effect of deregulation on optimal price-quality rivalry, entry auctions and strategic behavior under cross-market price restrictions, capital budgeting in a de-centralized firm under uncertainty, Bayesian estimates of market risk, measuring investor growth expectations, and cash management. Dr. Vander Weide's research has been published in the *American Economic Review*, *Financial Management*, *International Journal of Industrial Organization*, *Journal of Finance*, *Journal of Financial and Quantitative Analysis*, *Journal of Bank Research*, *Journal of Portfolio Management*, *Journal of Accounting Research*, *Journal of Cash Management*, *Management Science*, *Atlantic Economic Journal*, *Journal of Economics* *and Business*, and *Computers and Operations*. Dr. Vander Weide authored a book entitled *Managing Corporate Liquidity: An Introduction to Working Capital Management* published by John Wiley and Sons, Inc. Dr. Vander Weide is also a founder and president of Financial Strategy Associates, a consulting firm that provides strategic, financial, and economic consulting services to corporate clients, including expert witness testimony. As an expert witness, Dr. Vander Weide has testified in more than 400 proceedings before state and federal agencies and courts on such topics as economic and competitive conditions in the telecommunications, electric, and gas industries, cost of capital, mergers and acquisitions, incentive regulation, economic pricing guidelines, and valuation.

Timothy W. Viezer is the Managing Director – Investments at the University of Cincinnati where he assists the Chief Investment Officer in managing the University's endowment. Previously, Dr. Viezer was the Senior Investment Officer – Equities at the School Employees Retirement System of Ohio (SERS) where he oversaw 30 external managers in SERS' $7 billion domestic and international equity portfolios. Before joining SERS, Dr. Viezer worked at the State Teachers Retirement System of Ohio as a real estate portfolio analyst and at the Centerior Energy Corporation as a corporate economist. Dr. Viezer received his BBA and MAE in economics and his BA and MA in political science at Kent State University and he earned his Ph.D. in agricultural and applied economics at The Ohio State University. Dr. Viezer is a CFA charterholder and a Chartered Alternative Investment Analyst (CAIA) designee. He has served as the President of the Columbus Association of Business Economists and as the Vice President of the CFA Society of Columbus. Dr. Viezer has published several articles on real estate portfolio management and has taught over 15 years as an adjunct professor in several colleges and universities.

Ganlin Xu is the Vice President of Research & Development at GuidedChoice.com, Inc. After receiving a Ph.D. in Applied Mathematics from Carnegie Mellon University, he joined Dr. Markowitz's global portfolio research group at Daiwa Securities. He is interested in various theoretical and practical aspects of portfolio management, and has published in, among numerous others, the *SIAM Journal of Control and Optimization*. Since 2000, Dr. Xu has lead an effort at GuidedChoice incorporating Dr. Markowitz's mean variance efficient frontier into innovative products that provide investment advice and managed accounts to millions of investors.

William T. Ziemba is the Alumni Professor (Emeritus) of Financial Modeling and Stochastic Optimization in the Sauder School of Business, University of British Columbia where he taught from 1968 to 2006. His Ph.D. is from the University of California, Berkeley. He has been a visiting professor at Cambridge, Oxford, London School of Economics, University of Reading (ICMA Centre), and Warwick in the UK, at Stanford, UCLA, Berkeley, MIT, University of Washington and Chicago in the USA, Universities of Bergamo, Venice and Luiss in Italy, Tsukuba in Japan, and the National University of Singapore. He has been a consultant to a number of leading financial institutions including the Frank Russell Company, Morgan Stanley, Buchanan Partners, RAB Hedge Funds, Gordon Capital, and Private International Wealth Management. His research is in asset-liability management, portfolio theory and practice, security market imperfections, Japanese and Asian financial markets, sports and lottery investments, and applied stochastic programming. He has published widely in journals such as *Operations Research*, *Management Science*, *Mathematics of OR*, *Mathematical Programming*, *American Economic Review*, *Journal of Finance*, and *JFQA* and in many books and special journal issues. Recent books include *Applications of Stochastic Programming*, with S.W. Wallace, SIAM-MPS, 2005,*Stochastic Optimization Models in Finance*, 2nd edition with R.G. Vickson, World Scientific, 2006 and *Handbook of Asset and Liability Modeling*, Volume 1: *Theory and Methodology* (2006) and Volume 2: *Applications and Case Studies* (2007) with S.A. Zenios, North Holland, *Scenarios for Risk Management and Global Investment Strategies* with Rachel Ziemba, Wiley, 2007 and *Handbook of Investments: Sports and Lottery Betting Markets*, with Donald Hausch, North Holland, 2008. His co-written practitioner paper on the Russell-Yasuda model won second prize in the 1993 Edelman Practice of Management Science Competition. He has been a futures and equity trader and hedge fund and investment manager since 1983. He is the series editor for North Holland's *Handbooks in Finance*, and previously was the CORS editor of INFOR and the department of finance editor of *Management Science*, 1982–1992.

Index

F