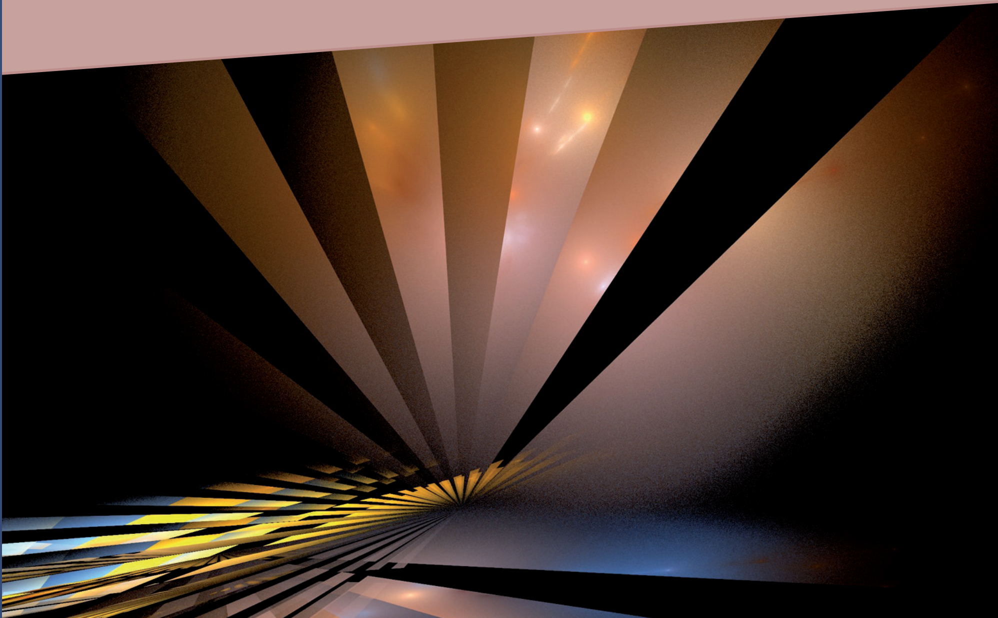
Risk
books

Portfolio Construction and Risk Budgeting

4th Edition

BY BERND SCHERER



Portfolio Construction and Risk Budgeting

Fourth Edition

Portfolio Construction and Risk Budgeting

Fourth Edition

Bernd Scherer

Risk
books

Published by Risk Books, a Division of Incisive Financial Publishing Ltd

Haymarket House
28–29 Haymarket
London SW1Y 4RX
Tel: + 44 (0)20 7484 9700
Fax: + 44 (0)20 7484 9797
E-mail: books@incisivemedia.com
Sites: www.riskbooks.com
www.incisivemedia.com

ISBN 978-1-906348-35-9

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

Publisher: Nick Carver
Commissioning Editor: Sarah Hastings
Managing Editor: Jennifer Gibb
Typeset by T&T Productions Ltd, London

Printed and bound in the UK by PrintonDemand-Worldwide

Every effort has been made to ensure the accuracy of the text at the time of publication, this includes efforts to contact each author to ensure the accuracy of their details at publication is correct. However, no responsibility for loss occasioned to any person acting or refraining from acting as a result of the material contained in this publication will be accepted by the copyright owner, the editor, the authors or Incisive Media.

Many of the product names contained in this publication are registered trade marks, and Risk Books has made every effort to print them with the capitalisation and punctuation used by the trademark owner. For reasons of textual clarity, it is not our house style to use symbols such as TM, ®, etc. However, the absence of such symbols should not be taken to indicate absence of trademark protection; anyone wishing to use product names in the public domain should first clear such use with the product owner.

While best efforts have been intended for the preparation of this book, neither the publisher, the editor nor any of the potentially implicitly affiliated organisations accept responsibility for any errors, mistakes and or omissions it may provide or for any losses howsoever arising from or in reliance upon its information, meanings and interpretations by any parties.

To Katja, Leonhard, Sebastian, Jana and Katharina

Contents

About the Author

Bernd Scherer teaches finance at EDHEC Business School and is a member of EDHEC Risk. He is also on the management board of the London Quant Group. Before joining EDHEC he was managing director and global head of Quantitative Asset Allocation at Morgan Stanley Investment Management, where he was responsible for the creation of active investment strategies within commodities, foreign exchange, credit and volatility products. During his 16-year career in asset management he has held various senior positions at Morgan Stanley, Deutsche Bank, Oppenheim Investment Management and J. P. Morgan Investment Management. Bernd's current research interests focus on signal construction, portfolio optimisation and asset liability modelling. He has written six books and more than 50 publications in leading academic and practitioner journals, such as the *Journal of Banking and Finance*, *Journal of Financial Markets*, *Journal of Economics and Statistics*, *Journal of Portfolio Management*, *Financial Analysts Journal*, *Journal of Investment Management*, *Risk*, *Financial Markets and Portfolio Management*, *Journal of Asset Management*, etc. Bernd holds MBA and MSc degrees from the University of Augsburg and the University of London, as well as a PhD in finance from the University of Giessen.

Introduction

OBJECTIVES AND CONTENT

The fourth edition of this highly successful book provides a comprehensive and up-to-date treatment of alternative portfolio construction techniques, ranging from traditional methods based on mean–variance and lower partial moments approaches, through Bayesian techniques, to more recent developments such as portfolio resampling and stochastic programming solutions using scenario optimisation. Most chapters have been considerably extended to cover the rapid expansion of the literature since the third edition was published. Most chapters now also include exercise sections.

Chapter 1 starts with a review of Markowitz-based solutions, with a particular focus on multifund separation of the role of characteristic portfolios and active management decisions. It also questions the suitability of mean–variance investing as currently applied to practical investment problems.

Chapter 2 extends this analysis into non-standard applications, such as the use of cluster analysis to redefine the investment universe, the treatment of illiquid asset classes, life-cycle investing or time-varying covariance matrices.

Chapter 3 moves away from the classical model and introduces non-normality. It provides a toolkit that will enable the user to judge when non-normality is a problem and when it is not. Lower-partial-moments-based portfolio construction is carefully discussed, and the chapter covers all the mathematical tools needed to apply this important technique to real-world portfolio problems.

Chapter 4 introduces estimation error and shows how to deal with it heuristically using portfolio resampling or constrained optimisation. Particular attention is given to the concept of resampled efficiency, which has recently appeared in the literature, as an increasing number of investors become interested in this approach to estimation error.

Chapter 5 provides a critical overview on robust optimisation and uses a common data set to apply robust optimisation techniques to portfolio construction exercises. New focus is given to norm constraints on optimal portfolio weights to directly enforce diversification.

Chapter 6 deals with estimation error from a more conventional angle, reviewing various Bayesian techniques. Close attention is paid to problems with data, particularly time series of different lengths and how to deal them with, since they represent one of the main data problems faced by practitioners. This chapter now also includes the use of hierarchical priors in manager allocation as well as recent advances in the Black–Litterman framework.

Chapter 7 introduces out-of-sample testing of portfolio construction methodologies. Following our detailed critique of resampled efficiency in Chapter 4 a particular emphasis has been given to resampled efficiency and its alternatives.

Chapter 8 deals with the most overlooked problem in practical portfolio construction: transaction costs. We show how various forms of transaction costs can be incorporated into the portfolio construction process. A new section on optimal trade scheduling has also been added.

Chapter 9 introduces options into the portfolio construction process. This subject has not been treated at all in textbooks and the reader will learn how to calibrate and use both the risk neutral and the real-world distribution in portfolio construction examples.

Chapter 10, on scenario optimisation, is a natural extension of the four previous chapters. It describes the most general form of portfolio optimisation that can simultaneously deal with data problems (estimation error, time series of different lengths) and with non-linear instruments, non-normal distributions and non-standard preferences.

Chapter 11 describes the process of budgeting active manager risk. It provides the mathematical tools to address such questions, such as how much of a fund to have actively managed, where to be active, and whether core–satellite investing is superior to enhanced indexing.

Chapter 12 deals explicitly with key concepts in making benchmark-relative decisions. Again, the focus is on problems that are rarely handled in traditional textbooks, such as implicit funding assumptions and risk decomposition, optimisation against multiple benchmarks, and tracking error and its forecasting ability. The latter is considered through a comparison of the efficiency of the tracking error approach with that of mean–variance methods.

Chapter 13 provides insights into the construction of realistic long–short portfolios as well as into the impact of constraints on value added. A new methodology is introduced that differs from the well-known transfer coefficient and is able to trace implementation costs down to the security level.

Chapter 14 focuses on the optimal tracking error choice from the asset manager's perspective. This is a change from previous chapters that have been looking at the asset management world from a client's point of view. It derives conditions under which excessive risk taking can be avoided, which is at the heart of the current regulatory debate.

Chapter 15 adds the time dimension to portfolio choice. How do optimal portfolios change if we introduce predictability in asset returns? This chapter will reveal that predictability leads to horizon-dependent asset allocation as well demand for tactical variations.

Chapter 16 looks at inherent risks in the asset management business. This is a break with the current literature, which is consumed with managing clients' risks rather than its own risks. Topics like managing market risks in asset-based fees or managing the risk of fund outflows are covered in great detail.

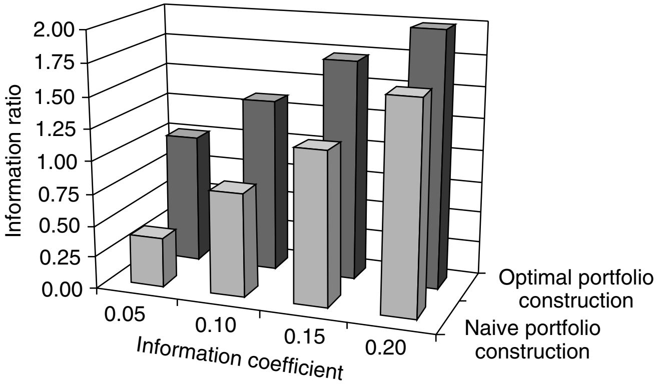
TARGET AUDIENCE

Anyone developing, managing or selling financial products – whether on the buy or the sell side – needs to know how his or her financial products fit into an investor's portfolio and how much the investor might find optimal. Hence, an understanding of how a scarce risk budget is optimally allocated is essential.

This book has been written primarily for practitioners, including portfolio managers, consultants, strategists, marketers and quantitative analysts. It is largely self-contained, and so could also prove useful to final-year undergraduates and MBAs who wish to extend their knowledge beyond the narrow world of mean–variance-based solutions typically taught at business schools.

Great care has been taken to illustrate theoretical concepts with simple examples that can be reproduced by readers to check their understanding of the methodology involved. This should allow practitioners to apply the algorithms discussed in the text to everyday problems with the minimum difficulty. The book is intended to be accessible to the mathematically interested practitioner who has a

Figure 1 Comparison of results of naive and optimal approaches to portfolio construction



Height of bars represents investment success for a given level of investment skill (skill measured as information coefficient).

basic understanding of calculus, matrix algebra and statistics. Some knowledge of financial economics will also prove helpful.

We think the book comes at the right time as it reviews all the major portfolio construction techniques at a time when portfolio construction is on everybody's radar screen.[1]

IMPORTANCE OF PORTFOLIO CONSTRUCTION

Portfolio construction – meaning the optimal implementation of a set of "signals" generated by strategists, asset allocators, analysts and the like – is at the centre of any modern investment process.[2] This has not always been the case, and cynics might argue that the increased interest in risk management techniques in the asset management arena is in large part a response to the significant underperformance of many investment houses along with its unpleasant consequences, such as legal actions, compensation and/or a shift to passive management. Although that might not apply in all cases, the asset management industry as a whole is undergoing a more realistic assessment of its performance-generating skills.

At this point it is worth drawing the reader's attention to a finding that goes some way to explaining the asset management industry's focus on portfolio construction. For different skill levels we compared the outcome of naive portfolio construction (equally overweight/underweight assets with the best/worst signals) with optimal portfolio construction (optimally over/underweight assets depending on signal strength as well as risk contributions). Skill levels were measured as an information coefficient, ie, the correlation between forecast returns and actual returns. Performance was measured as the information ratio, which is the active return divided by the active risk, ie, tracking error. The results are summarised in Figure 1.[3]

At realistic levels of investment skill (information coefficient of 0.05), optimal portfolio construction does make a difference, giving an information ratio about four times higher than that of the naive method. As the skill level rises (moving to the right in the graph), the gap narrows considerably, confirming the intuition that you do not need risk management if you have good forecasts. However, investment skill has to be tripled, from 0.05 to 0.15, to arrive at the same risk-adjusted performance as optimal portfolio construction. Given that an information coefficient of 0.05 might be viewed as the level achieved by the average asset manager, Figure 1 suggests there is good reason for the increasing focus on portfolio construction methods with a sound analytical basis.

PORTFOLIO CONSTRUCTION VERSUS RISK BUDGETING

Contrary to the belief of some, there is no difference between portfolio construction using various portfolio optimisation tools that attempt to trade off expected return against expected risk (return risk, estimation risk, modelling risk, etc) and risk budgeting. That is why this book has been titled "portfolio construction and risk budgeting". Investors have to trade off risk and return in an optimal way, ie, in a way that is both in accordance with their beliefs and preferences and which also ultimately optimises their portfolios. Practitioners may be confused as the result of the optimal portfolio construction exercise is an allocation either in nominal dollar terms or in percentage weights, whereas risk budgeting does not arrive at asset weights but at risk exposures expressed in terms of value-at-risk or percentage contributions to risk.

However, if risk budgets are optimally derived, this is just a presentational difference – one that certainly has educational value but with no investment value. Hence, the biggest advantage of the risk budgeting approach is that it becomes evident to investors that even small positions can carry large risks (a 5% departure from the benchmark allocation into emerging market debt for a portfolio benchmarked against government bonds will consume most of an investor's risk budget). The fact that portfolio optimisation reports results in terms of actual asset weights does not mean that a portfolio optimiser does not look at the risk contributions of different asset classes. Those who see risk budgeting simply as a way of enforcing diversification, and who diversify by assuming rather than deriving a solution, will fail – by over- or underdiversifying – to achieve the highest return per unit of risk and will therefore expose an investor to more risk than necessary.

ACKNOWLEDGEMENTS

I am indebted to Dean Barr, Christoph Bernard, Kenneth Yip, Greg van Invegen, Colm O'Cinneide, Ivillina Popova, Christopher Donohue, Ken Winston, Steve Satchell, Frank-Peter Martin, Dan DiBartolomeo, Jason McQueen and Lionel Martellini for invaluable discussions. All errors, of course, remain mine. I also want to thank Sarah Jenkins, who took on the first edition for Risk Books, my current editor, Sarah Hastings, and to Stan Beckers, who acted as a referee.

During the time between the first and fourth editions, I had the pleasure to work for Deutsche Bank and Morgan Stanley. I am greatly indebted to both organisations for their support and professionalism, and for their wealth of knowledge, some of which undoubtedly found its way into this book.

Writing a book like this at the same time as holding a challenging job also needs the support of your editor. I want to thank the staff at Risk Books, and, in particular, Jon Wainwright of T&T Productions Ltd, who has been extremely helpful, knowledgable and professional in bringing the fourth edition into a new format and ironing out previous glitches.

[1] The author appreciates comments and ideas and can be contacted at drberndscherer@gmx.net.

[2] Although the terms "investment process" and "investment philosophy" are often used interchangeably, we believe that there is an important distinction to be made. An investment *philosophy* comprises, among other things, an investor's beliefs concerning how markets operate, where inefficiencies arise and how they can be exploited; in short, the value proposition. The investment process refers to the organisational structure and defines best practice on how

to implement an existing investment philosophy most effectively. Portfolio construction and risk budgeting are, therefore, integral parts of an investment process but not of an investment philosophy. Investment philosophy is about signal generation, whereas the investment process concerns how a given set of signals is put to use.

[3] Chapter 8 provides a more detailed explanation of the calculations behind Figure 1.

1

A Primer on Portfolio Theory

1.1 MEAN–VARIANCE-BASED PORTFOLIO CONSTRUCTION

The theory of mean–variance-based portfolio selection is a cornerstone of modern asset management.[1] It rests on the presumption that rational investors choose among risky assets purely on the basis of expected return and risk, with risk measured as variance. In this case a portfolio is considered mean–variance-efficient if it minimises the variance for a given expected mean return or if it maximises the expected mean return for a given variance. Mean–variance efficiency rests on firm theoretical grounds if either:

- investors exhibit quadratic utility – in which case they ignore non-normality in the data[2]; or
- returns are multivariate normal – in which case the utility function is irrelevant as all higher moments, such as skewness or kurtosis, can be expressed as a function of mean and variance and, hence, all optimal solutions satisfy the mean–variance criterion.[3]

Both assumptions will be relaxed in Chapters 3–10.

1.1.1 Mean–variance optimisation in an asset-only world

We start with the solution to the portfolio construction problem in a world without stochastic liabilities, ie, where liabilities come in the form of a fixed-hurdle rate (like cash).

Suppose we know the $k \times 1$ vector of expected returns, $\boldsymbol{\mu}$, where the ith element represents the expected return over cash, c,[4] and that we also know the $k \times k$ covariance matrix of returns, $\boldsymbol{\Omega}$.[5] We now want to find the $k \times 1$ vector of optimal portfolio weights, $\boldsymbol{w}^*$. The matrix and the two vectors can be written as

$$\boldsymbol{\Omega} = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_{k1} & \cdots & \sigma_{kk} \end{bmatrix}, \qquad \boldsymbol{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_k \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_k \end{bmatrix} \tag{1.1}$$

where $\sigma_{ij} = \operatorname{cov}(R_i - c, R_j - c)$ and $\mu_i = E(R_i - c)$, where R_i is the total return of the ith asset. Portfolio risk, σ_{p}^2, measured as variance, and portfolio return, μ_{p}, are calculated from[6]

$$\sigma_{\mathrm{p}} = \begin{bmatrix} w_1 \\ \vdots \\ w_k \end{bmatrix}' \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_{k1} & \cdots & \sigma_{kk} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_k \end{bmatrix}, \qquad \mu_{\mathrm{p}} = \begin{bmatrix} w_1 \\ \vdots \\ w_k \end{bmatrix}' \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_k \end{bmatrix} \tag{1.2}$$

where primes indicate a transposed matrix. In practice, we can either:

- minimise portfolio variance for all portfolios ranging from minimum return to maximum return to trace out an efficient frontier (the geometric location of all mean–variance-efficient portfolios, ie, of future investment opportunities); or
- construct optimal portfolios for different risk-tolerance parameters, λ, and, by varying λ, find the efficient frontier.

We follow the latter approach, which trades off risk against return by maximising

$$\text{Utility} \approx \mu_{\mathrm{p}} - \frac{1}{2\lambda}\sigma_{\mathrm{p}}^2 = \boldsymbol{w}'\boldsymbol{\mu} - \frac{1}{2\lambda}\boldsymbol{w}'\boldsymbol{\Omega}\boldsymbol{w} \tag{1.3}$$

for various risk-tolerance parameters. “Utility” is a measure of happiness. The higher our risk tolerance, the less weight is given to the variance (penalty) term and the more aggressive our portfolios will become.[7]

The optimal solution[8] is found by taking the first derivative with respect to portfolio weights[9], setting the term to zero and solving for the optimal weight vector, $\boldsymbol{w}^*$

$$\left.\begin{aligned} \frac{\mathrm{d}\,\text{Utility}}{\mathrm{d}\boldsymbol{w}} &= \boldsymbol{\mu} - \frac{1}{2\lambda}2\boldsymbol{\Omega}\boldsymbol{w} = \boldsymbol{\mu} - \frac{1}{\lambda}\boldsymbol{\Omega}\boldsymbol{w} = \boldsymbol{0} \\ \boldsymbol{w}^* &= \lambda\boldsymbol{\Omega}^{-1}\boldsymbol{\mu} \end{aligned}\right\} \tag{1.4}$$

Note that the portfolio in Equation 1.4 does not add up to 100%, but we can think of cash as the difference remaining because cash has a zero risk premium and no variance. The optimal allocation into risky assets rises with higher risk tolerance, higher return and diminishing uncertainty about the mean return (the inverse of a covariance matrix is a measure of the precision of mean returns).

Let us see whether we can use this basic insight for a practical application. Equation 1.4 implies that investors with different risk tolerance would hold very similar portfolios, ie, portfolios would only differ by a multiple λ. In short they only differ by leverage (alternatively, you can say by the level of cash holdings) if you want to use a more neutral wording. Now suppose you see the following asset allocation advice. A risk-averse investor (A) is recommended to invest into 60% cash / 20% bonds / 20% equities while the same asset manager advises a more conservative investor (B) to invest into 40% cash / 10% bonds / 50% equities. Can you see the inconsistency? While we would expect that a more conservative investor B would hold more cash, we would not expect that the relative weight between equities and bonds changes. In fact Equation 1.4 told us it would not as portfolios would only differ by leverage (split between cash and risky assets) but the relative proportion between risky assets would remain untouched.

What would happen if we had no return estimates (or if all return estimates were the same)? What criteria could we then use for portfolio construction? The answer to both questions is the minimum risk portfolio. We find it by solving

$$\boldsymbol{w} = \arg\min_{\boldsymbol{w}} \tfrac{1}{2}\boldsymbol{w}'\boldsymbol{\Omega}\boldsymbol{w} - \upsilon(\boldsymbol{w}'\mathbf{1} - 1) \tag{1.5}$$

Taking derivatives with respect to both weight vector, $\boldsymbol{w}$, and Lagrange multiplier, υ, we get

$$L_w = \boldsymbol{\Omega}\boldsymbol{w} - \upsilon\mathbf{1} = 0 \tag{1.6}$$

$$L_\upsilon = \boldsymbol{w}'\mathbf{1} - 1 = 0 \tag{1.7}$$

Now we will undertake the following steps. Solve the first equation for the weight vector and substitute the result into the second equation to arrive at an expression for the Lagrange multiplier. Take the Lagrange multiplier and substitute back into the first equation above and we have found the minimum variance portfolio. For reasons to become clear later we will denote it as $\boldsymbol{w}_1$. Our calculations should read as follows

$$\boldsymbol{w}_1 = \boldsymbol{\Omega}^{-1}\mathbf{1}\upsilon \tag{1.8}$$

$$(\boldsymbol{\Omega}^{-1}\mathbf{1}\upsilon)'\mathbf{1} - 1 = 0 \quad \Rightarrow \quad \upsilon = (\mathbf{1}'\boldsymbol{\Omega}^{-1}\mathbf{1})^{-1} \tag{1.9}$$

$$\boldsymbol{w}_1 = (\mathbf{1}'\boldsymbol{\Omega}^{-1}\mathbf{1})^{-1}\boldsymbol{\Omega}^{-1}\mathbf{1} \tag{1.10}$$

$$\begin{aligned}\sigma_1 &= \boldsymbol{w}_1'\boldsymbol{\Omega}\boldsymbol{w}_1 = (\boldsymbol{\Omega}^{-1}\mathbf{1}\upsilon)'\boldsymbol{\Omega}(\boldsymbol{\Omega}^{-1}\mathbf{1}\upsilon)\\ &= \mathbf{1}'\boldsymbol{\Omega}^{-1}\mathbf{1}\upsilon^2 = \frac{1}{\mathbf{1}'\boldsymbol{\Omega}^{-1}\mathbf{1}} \end{aligned}\tag{1.11}$$

Why is this important? Our interest in the minimum variance portfolio is focused around two areas of interest. On a theoretical basis it can be shown that many portfolios are subject to two-fund separation, ie, optimal portfolios can be shown to consist of a combination of minimum variance portfolio and speculative demand (see Chapter 5 on robust portfolio optimisation and estimation error). Also the minimum variance portfolio allows us to separate the effects of estimation error in risk estimates from estimation error in return estimates (see Chapter 6 on Bayesian analysis and portfolio choice). For practitioners recent evidence has shown that the minimum variance portfolio often outperforms market capitalised portfolios. This is clearly at odds with prediction from the CAPM that all combinations of cash and a well-diversified market portfolio dominate the minimum variance portfolio (offer a higher return per unit of risk). It seems that these results are mainly due to the "high beta anomaly". High beta stocks earn less than what the CAPM predicts they should earn in equilibrium.[10] The minimum variance portfolio will by construction show more exposure to low beta stocks. This point is easy to proof. Beta for the minimum variance portfolio is given by

$$\beta_1 = \frac{\boldsymbol{w}_1' \boldsymbol{\Omega} \boldsymbol{w}_{\mathrm{m}}}{\boldsymbol{w}_{\mathrm{m}}' \boldsymbol{\Omega} \boldsymbol{w}_{\mathrm{m}}} = \frac{1}{\sigma_{\mathrm{m}}^2} \left(\frac{\boldsymbol{\Omega}^{-1}\mathbf{1}}{\mathbf{1}\boldsymbol{\Omega}^{-1}\mathbf{1}} \right)' \boldsymbol{\Omega} \boldsymbol{w}_{\mathrm{m}} \tag{1.12}$$

Expanding Equation 1.12 we get

$$\begin{aligned} \frac{1}{\sigma_{\mathrm{m}}^2} \frac{\mathbf{1}\boldsymbol{\Omega}^{-1}}{\mathbf{1}'\boldsymbol{\Omega}^{-1}\mathbf{1}} \boldsymbol{\Omega} \boldsymbol{w}_{\mathrm{m}} &= \frac{1}{\sigma_{\mathrm{m}}^2} \frac{1}{\mathbf{1}'\boldsymbol{\Omega}^{-1}\mathbf{1}} \mathbf{1}' \boldsymbol{w}_{\mathrm{m}} \\ &= \frac{1}{\sigma_{\mathrm{m}}^2} \frac{1}{\mathbf{1}'\boldsymbol{\Omega}^{-1}\mathbf{1}} \\ &= \frac{\sigma_1^2}{\sigma_{\mathrm{m}}^2} \\ &< 1 \end{aligned} \tag{1.13}$$

where we used that $\mathbf{1}'\boldsymbol{w}_{\mathrm{m}} = 1$, ie, the market portfolio sums to one, and $1/(\mathbf{1}'\boldsymbol{\Omega}^{-1}\mathbf{1}) = \sigma_1^2$. The minimum variance portfolio will have a beta smaller than one by construction.

1.1.2 Diversification

The idea of diversification is strongly connected with portfolio theory. It is therefore important to understand it well and address some of the most common errors. First, diversification arises from subdividing risks, ie, from allocating fractions of a given wealth to different assets. Adding on assets will not diversify (decrease) risks, but increase them. If you own 1 million European equities and you add another million US equities, your total risk has increased. If this sounds too obvious let us investigate two less obvious applications.

- A pension fund runs an almost matching fixed-income portfolio against its liabilities that also contain some actuarial noise (movements in the value of liabilities that cannot be hedged using capital market investments). A consultant suggests that you should take "alpha" risks in your assets as they diversify away when combined with the uncorrelated actuarial noise. Wrong! This will always increase risk not decrease it, because you add on risk and not subdivide it.

- A consultant claims that if asset returns are uncorrelated over time, this means you can diversify risks over time. This is wrong as you are adding not subdividing risks. Each investment period you again put your wealth at stake. Repeating this many times must be more risky than just doing it once.

The second question we need to ask ourselves is: is diversification a good thing? Well, not necessarily. First, if all your assets are so different that they share no common correlation, portfolio risk will be diversifiable and you should expect the risk-free rate, rather than a premium. Second, diversification always needs to be traded off against risk, a point we will return at this book at great length. Very diversified portfolios are unlikely to be efficient (offer the highest return per unit of risk). Unless you are extremely pessimistic about the quality of your inputs, you do not want to overdiversify.

1.1.3 Building blocks of portfolio choice: characteristic portfolios

In section 1.1.1 we solved our first portfolio optimisation problem by trading off returns versus risks. However, portfolio construction can be much more generic like this. In fact, what are the building blocks of optimal portfolios? We could, for example, desire to create a portfolio with "unit" exposure to a characteristic (an attribute of a particular asset). We might want to get exposure to characteristic linked to valuation (price to book, etc), size (log of market capitalisation), sentiment (earning upgrades to downgrade) or quality (accruals), to name a few. Is there a combination of "elementary portfolios" that we could use?

Even though all of these characteristics are not expected returns, it is straightforward to show that minimising portfolio risk subject to a unitary exposure to a characteristic, $\boldsymbol{\xi}$ ($k \times 1$ vector of security attributes) can be expressed as

$$\min_{w} \tfrac{1}{2}\boldsymbol{w}'\boldsymbol{\Omega}\boldsymbol{w} - \upsilon(\boldsymbol{w}'\boldsymbol{\xi} - 1) \tag{1.14}$$

with the solution to Equation 1.14 given by

$$\boldsymbol{w}_{\boldsymbol{\zeta}} = (\boldsymbol{\xi}'\boldsymbol{\Omega}^{-1}\boldsymbol{\xi})^{-1}\boldsymbol{\Omega}^{-1}\boldsymbol{\xi} \tag{1.15}$$

A careful look at Equation 1.10 reveals that the minimum variance portfolio is also a characteristic portfolio ($\boldsymbol{\xi} = \mathbf{1}$). Other prominent characteristic portfolios are the unit beta (asset betas as characteristics) or maximum Sharpe ratio (expected excess returns as characteristics) portfolio.

Suppose we want to build a portfolio that provides an investor with the highest expected return but that is also beta neutral (zero beta with respect to a market proxy) and cash neutral (self-financing long–short portfolio). The quickest way to do this is to follow Sorensen *et al* (2007) and conjecture that the optimal solution arises from three-fund separation, ie, as the combination of three characteristic portfolios with a different definition for $\boldsymbol{\xi}$

$$\boldsymbol{w} = \theta_{\boldsymbol{\mu}}\boldsymbol{w}_{\boldsymbol{\mu}} + \theta_1\boldsymbol{w}_1 + \theta_{\boldsymbol{\beta}}\boldsymbol{w}_{\boldsymbol{\beta}} \tag{1.16}$$

We get a system of three equations, ie, linear combination of characteristic portfolio that generates a predefined exposure

$$\theta_{\boldsymbol{\mu}}\boldsymbol{w}_{\boldsymbol{\mu}}'\boldsymbol{\mu} + \theta_1\boldsymbol{w}_1'\boldsymbol{\mu} + \theta_{\boldsymbol{\beta}}\boldsymbol{w}_{\boldsymbol{\beta}}'\boldsymbol{\mu} = 1 \tag{1.17}$$

Equation 1.17, for example, calculates the portfolio return for the sum of all three characteristic portfolios as the combination of the maximum Sharpe ratio portfolio return (characteristic exposure of one as $\boldsymbol{w}_{\boldsymbol{\mu}}'\boldsymbol{\mu} = 1$ by definition) and the returns of characteristic portfolios for "beta" and "one". Solving the system for the weights of the characteristic portfolios ($\theta_{\boldsymbol{\mu}}, \theta_1, \theta_{\boldsymbol{\beta}}$) will generate a portfolio with unit exposure to expected returns that is both cash as well as beta neutral. Of course, it is not always guaranteed that a unique solution exists. It will depend on the rank of the "coefficient matrix" in

$$\begin{bmatrix} \theta_{\boldsymbol{\mu}} \\ \theta_1 \\ \theta_{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} 1 & \boldsymbol{w}_1'\boldsymbol{\mu} & \boldsymbol{w}_{\boldsymbol{\beta}}'\boldsymbol{\mu} \\ \boldsymbol{w}_{\boldsymbol{\mu}}'\mathbf{1} & 1 & \boldsymbol{w}_{\boldsymbol{\beta}}'\mathbf{1} \\ \boldsymbol{w}_{\boldsymbol{\mu}}'\boldsymbol{\beta} & \boldsymbol{w}_1'\boldsymbol{\beta} & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \tag{1.18}$$

Characteristic portfolios have their origin in asset pricing as well as the development of quantitative investment strategies. After all, researchers in these areas want to calculate the payout from a given stock characteristic (profitability measure, etc) to see whether the characteristic is a priced factor (has an excess return that is not already explained by existing risk factors). Generally, these researchers want to find characteristics that can explain the cross section of stock returns. For each point in time they estimate

$$\boldsymbol{r} = \boldsymbol{\xi} r_{\boldsymbol{\xi}} + \boldsymbol{\varepsilon} \tag{1.19}$$

where $\boldsymbol{r}$ denotes a vector of cross-sectional excess returns (for example, one excess return for each of the constituents of the S&P 500 in September 2008, ie, a 500×1 vector), $\boldsymbol{\xi}$ denotes a 500×1 vector of characteristics (sometimes called factors) for these stocks (for example, leverage) at the beginning of September 2008 and $r_{\boldsymbol{\xi}}$ is our unknown regression parameter in this regression. We can interpret $r_{\boldsymbol{\xi}}$ as the factor return for leverage in September 2008. Repeating many cross-sectional regressions for a given time period (let us say the last 10 years) will yield a 120 leverage factor return realisations and we can apply standard tests for significance. Note that we did not control for other variables in Equation 1.19 so we cannot conclude we found a priced factor based on a significance test.[11] One way out of this is to use model-adjusted excess returns (residuals of multifactor asset pricing model) instead of raw returns in Equation 1.19. In any case the GLS estimate for the factor return for a given period is given by

$$r_{\boldsymbol{\xi}} = (\boldsymbol{\xi}'\boldsymbol{\Omega}^{-1}\boldsymbol{\xi})^{-1}\boldsymbol{\xi}'\boldsymbol{\Omega}^{-1}\boldsymbol{r} = \boldsymbol{w}_{\boldsymbol{\xi}}'\boldsymbol{r} \tag{1.20}$$

where $\boldsymbol{\Omega}$ denotes the covariance of asset returns. Note that $\boldsymbol{\Omega}$ is not generally known in a single cross-sectional regression, but we will ignore this technicality here as it can be efficiently dealt with). The important point here is that our characteristic portfolio Equation 1.15 multiplied with the cross section of portfolio returns creates a factor return that is identical to the factor return from a cross-sectional GLS regression.

1.1.4 Heuristic practitioner rules: portfolio ranks and sorts

Portfolio optimisation is a formal subject that requires investment in specific human capital as well as specialised software packages. At the same time practitioners have been deeply suspicious about a black box methodology where they could no longer necessarily relate outputs (weights) to inputs (returns). A positive view on a given stock or bond would not necessarily translate into a positive weight as an optimisation algorithm would trade off diversification losses against relative return advantages. What are the practical issues concerned with Equation 1.15? First, we note that the vector $\boldsymbol{\xi}$ might contain many outliers that will create unreasonably optimistic (concentrated in too few assets) allocations. Practitioners have long tried to smooth the input vector $\boldsymbol{\xi}$ by truncating outliers or throwing them back to the middle. Recent advances in robust optimisation seem to *ex post* justify this (see chapter 5). Second, we know that the covariance matrix can create highly leveraged bets in correlated assets that only marginally differ in expected returns. Therefore, many practitioners ignore the covariance matrix altogether, reasoning that covariance estimates become very noisy when the investment universe becomes large. A very common approach to portfolio construction is therefore to sort characteristics according to their rank and build quintile- (also called wing-) portfolios. These portfolios simply take the top and bottom 20% (or any other quintile below 50%) of names according to the characteristic used as sorting criterion and invest in them equally weighted as a long–short portfolio. While a covariance matrix can be used to leverage up these portfolios to the desired risk level the relative weighting of assets is not affected by risk or correlation differences. It is implicitly assumed that the portfolio manager can only distinguish between good and bad investments. However, that does not mean practitioners do not make very strong assumptions. Suppose, for example, a portfolio manager that will first divide the universe in the best 20% (score of $+1$) and the worst 20% (score of -1) of stocks. Portfolio weights will be calculated as score divided by residual volatility

$$w_{i,t}^{\pm} = \begin{cases} +\dfrac{1}{\sigma_{i,t}} & \text{upper } 20\% \\ -\dfrac{1}{\sigma_{i,t}} & \text{lower } 20\% \end{cases} \tag{1.21}$$

Each position size is therefore an inverse function of volatility. Riskier assets get smaller positions while all "good assets" will be overweight, while all bad assets are underweight. The equivalent to Equation 1.15 is given by[12]

$$
\begin{aligned}
w &= \boldsymbol{\Omega}^{-1}\boldsymbol{\xi} \\
&= \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_i^2 & & \\ \vdots & & \ddots & \\ 0 & & & \sigma_n^2 \end{bmatrix}^{-1} \begin{bmatrix} -1\sigma_1 \\ +1\sigma_1 \\ \vdots \\ -1\sigma_1 \end{bmatrix} = \begin{bmatrix} -\dfrac{1}{\sigma_1} \\ +\dfrac{1}{\sigma_i} \\ \vdots \\ -\dfrac{1}{\sigma_n} \end{bmatrix}
\end{aligned}
\tag{1.22}
$$

While practitioners like shortcuts and aim for "robust" results, it should be noted that the underlying "assumptions" are not less extreme in any case. Recently, Almgren and Chriss (2006) suggested a way to build optimal portfolios based on the idea of replacing expected returns by ordering information. In other words, how do we get from a sort to a portfolio. Practically, this is very interesting as it is often the only way portfolio managers are able to express their opinions, ie, "I prefer A over B over C, etc". The authors show we can create "pseudo" expected returns (called "centroids") for complete single sorts (all assets are sorted according to a single characteristic) of k assets from

$$
\xi_i^{\mathrm{c}} = N^{-1}\left(\frac{k + 1 - \operatorname{rank}(i) - \theta}{n - 2\theta + 1}\right) \tag{1.23}
$$

where $\operatorname{rank}(i)$ is the rank order information of the ith stock, $N(\cdot)^{-1}$ is the inverse of the cumulative normal distribution and $\theta = 0.4424 - 0.1185n^{-0.21}$. Optimal portfolios can now be calculated from either equalising weights with "centroids"

$$
w_i = \xi_i^{\mathrm{c}} \tag{1.24}
$$

or building characteristic portfolios from "centroids" instead.

1.1.5 Asset–liability management and the surplus-efficient frontier

A conceptual problem of the analysis so far has been that in practice we cannot isolate assets from liabilities. Investors without liabilities do not need assets. Asset–liability management[13] becomes imperative for every investor who seeks to define their potential liabilities carefully.[14] It focuses on managing the difference between assets and liabilities, also called "surplus". The change in surplus depends directly on the returns of the asset portfolio, R_{p}, as well as the liability returns (percentage changes in the value of outstanding liabilities), R_{l}

$$\Delta\text{Surplus} = \text{Assets} \times R_{\text{p}} - \text{Liabilities} \times R_{\text{l}} \tag{1.25}$$

We will express surplus returns as change in surplus relative to assets[15]

$$\begin{aligned}\frac{\Delta\text{Surplus}}{\text{Assets}} &= R_{\text{p}} - \frac{\text{Liabilities}}{\text{Assets}} R_{\text{l}} \\ &= R_{\text{p}} - f\, R_{\text{l}}\end{aligned} \tag{1.26}$$

where f is the ratio of liabilities to assets. If we set $f = 1$ and $R_{\text{l}} = c$, we are back in a world without liabilities (alternatively we could think of liabilities as cash). Surplus volatility, $\sigma^2_{\text{surplus}}$, can now be incorporated into the well-known framework of Equation 1.2 by including a short position in liabilities[16]

$$\sigma^2_{\text{surplus}} = \begin{bmatrix} w_1 \\ \vdots \\ w_1 \\ -f \end{bmatrix}' \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1k} & \sigma_{1l} \\ \vdots & \ddots & \vdots & \vdots \\ \sigma_{k1} & \cdots & \sigma_{kk} & \sigma_{kl} \\ \sigma_{l1} & \cdots & \sigma_{lk} & \sigma_{ll} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_1 \\ -f \end{bmatrix} \tag{1.27}$$

We assume that liabilities can be summarised as one single asset, l, whereas, for example, σ_{kl} summarises the covariance of the kth asset with our liabilities.[17] If assets equal liabilities, we arrive at a traditional active optimisation where liabilities directly play the role of a benchmark asset. How, then, can we transform the asset–liability management problem into the well-known portfolio optimisation so that we need not change portfolio optimisation algorithms and can still search for the optimal solution in terms of assets only? All we have to do is to express the covariance matrix in terms of surplus risk, ie, as the volatility of return differences between assets and liabilities (very much like any other active optimisation problem). The covariance matrix of assets and liabilities is transformed via a matrix of long–short positions into the covariance matrix of surplus returns

$$\begin{aligned}\boldsymbol{\Omega}_{\text{surplus}} &= \begin{bmatrix} 1 & 0 & \cdots & 0 & -f \\ 0 & 1 & & & -f \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & 1 & -f \end{bmatrix} \\ &\quad \times \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1k} & \sigma_{11} \\ \vdots & \ddots & & \vdots \\ \sigma_{k1} & & \sigma_{kk} & \sigma_{kl} \\ \sigma_{l1} & \cdots & \sigma_{lk} & \sigma_{11} \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 & -f \\ 0 & 1 & & & -f \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & 1 & -f \end{bmatrix}' \end{aligned} \tag{1.28}$$

As well as offering the convenience of using the same portfolio optimisation algorithm, writing the covariance matrix in terms of surplus variance will lead to a much better understanding of risk in an asset–liability framework. While cash, for example, is a very conservative asset in an asset-only framework, in an asset–liability framework it becomes one of the most risky assets as it has no covariation with liabilities (often similar to long bonds) and therefore cannot serve as a liability-hedging asset.[18] To remain consistent, we manipulate the expected returns to reflect the relative return of assets versus liabilities

$$\boldsymbol{\mu}_{\text{surplus}} = \begin{bmatrix} \mu_1 - f\mu_l \\ \vdots \\ \mu_k - f\mu_l \end{bmatrix} + c(1-f) \tag{1.29}$$

As with Equation 1.61, we maximise

$$\text{Utility} \approx \mu_{\text{surplus}} - \frac{1}{2\lambda}\sigma_{\text{surplus}}$$

subject to $\boldsymbol{A}\boldsymbol{w} = \boldsymbol{b}$, which yields[19]

$$\begin{aligned}\boldsymbol{w}^* &= \boldsymbol{\Omega}^{-1}_{\text{surplus}}\boldsymbol{A}'(\boldsymbol{A}\boldsymbol{\Omega}^{-1}_{\text{surplus}}\boldsymbol{A}')^{-1}\boldsymbol{b} \\ &\quad + \lambda\boldsymbol{\Omega}^{-1}_{\text{surplus}}(\boldsymbol{\mu}_{\text{surplus}} - \boldsymbol{A}'(\boldsymbol{A}\boldsymbol{\Omega}^{-1}_{\text{surplus}}\boldsymbol{A}')^{-1}\boldsymbol{A}\boldsymbol{\Omega}^{-1}_{\text{surplus}}\boldsymbol{\mu}_{\text{surplus}})\end{aligned} \tag{1.30}$$

Varying the risk-tolerance parameter λ, we can trace out the surplus-efficient frontier (the geometric location of surplus-efficient portfolios. Unconstrained (asset-only) efficient frontier and surplus-efficient frontier coincide if

- liabilities are cash (or, equally, if assets have no covariation with liabilities);
- all assets have the same covariation with liabilities (it does not matter which asset is chosen);[20] or
- there exists a liability-mimicking asset and it lies on the efficient frontier (it is a general result in benchmark-relative optimisation that optimising relative to a mean–variance-inefficient benchmark (liabilities) will never give a mean–variance-efficient portfolio).[21]

We will illustrate the concepts outlined so far using a numerical example to show that asset-only and asset–liability solutions generally differ. Suppose that the funding ratio of a pension fund is one (assets equal liabilities) and that we have the following information[22]

$$\boldsymbol{\Omega} = \begin{bmatrix} 0.04 & 0.024 & 0.0028 \\ 0.024 & 0.0225 & 0.0021 \\ 0.0028 & 0.0021 & 0.0049 \end{bmatrix}$$

$$\boldsymbol{\Gamma} = \begin{bmatrix} 0.015 \\ 0.01125 \\ 0.00525 \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} 4\% \\ 3\% \\ 1.5\% \\ 2\% \end{bmatrix}, \qquad \sigma_{ll} = 0.0025$$

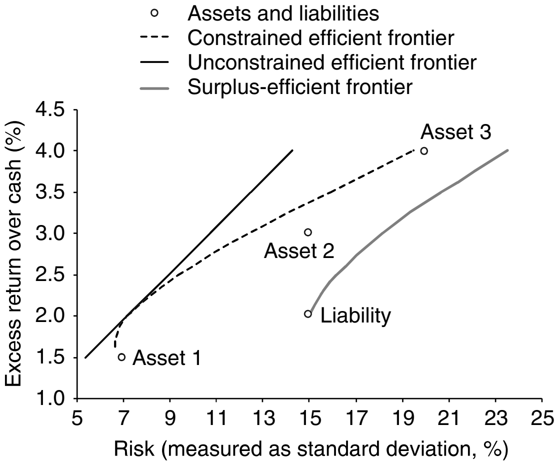
where $\boldsymbol{\Gamma}$ expresses the covariance between asset and liability returns. We can now directly calculate surplus volatility according to Equation 1.27

$$\sigma^2_{\text{surplus}} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ -1 \end{bmatrix}' \begin{bmatrix} \boldsymbol{\Omega} & \boldsymbol{\Gamma} \\ \boldsymbol{\Gamma}' & \sigma_{ll} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_1 \\ -1 \end{bmatrix}$$

Armed with these assumptions, we can now solve for both the unconstrained (Section 1.1.1) and the constrained (Section 1.6) efficient frontier as well as for the surplus-efficient frontier (Section 1.1.5) to confirm the points made above.

As shown in Figure 1.1, the surplus-efficient frontier is dominated by both (asset-) efficient frontiers, which offer a lower level of risk for the same level of expected return. This is not surprising as a different objective – optimising versus liabilities rather than cash – has been pursued; if we plotted the asset-efficient frontier in a

Figure 1.1 Efficient frontiers



surplus risk and return diagram, the opposite would be true. In a totally unconstrained optimisation the efficient frontier effectively becomes a straight line from the beginning as every portfolio can be leveraged without limit.[23] Asymptotically, the constrained frontiers also become nearly straight lines as the weights from the minimum-variance portfolio in Equation 1.65 have less and less influence.

We have seen in this section that asset-only and asset–liability management problems can be solved with the same analytical framework, applying essentially the same algorithms. All that is needed is a transformation of the original covariance matrix.

1.2 HOW WELL DOES MEAN–VARIANCE INVESTING WORK?

In reality, returns are not multivariate normal-distributed (see Chapter 3); investors do not exhibit quadratic utility and they do not live in a one-period world, ie, they have an investment horizon that lasts longer than one period and they are able to take corrective actions within that period.[24] In fact, they have multiperiod objectives and will revise their decisions at the end of each period. How serious are these limitations on the practical use of mean–variance-based portfolio construction? Here we will concentrate on three important questions.

1. How well does the mean–variance framework approximate reality, where investors might have different utility functions and returns might not be normally distributed?
2. How well does the one-period solution approximate multi-period optimality?
3. Who should use portfolio optimisation?

1.2.1 How well does the mean–variance framework approximate reality?

Suppose that an investor maximises a utility function, $u(R)$, that is a function of end-of-period wealth. Normalising start-of-period wealth to 1, we can write utility as a function of end-of-period return, R. We can now expand the utility function around this expected end-of-period wealth and take expectations to arrive at an approximation of expected utility, $E[u(R)]$[25]

$$E[u(R)] \approx \underbrace{u(\bar{R}) + \frac{1}{2}\frac{\mathrm{d}^2u(\bar{R})}{\mathrm{d}R^2}\sigma^2}_{\text{Quadratic approximation}} + \underbrace{\sum_{n=3}^{\infty}\frac{1}{n!}\frac{\mathrm{d}^nu(\bar{R})}{\mathrm{d}R^n}}_{\text{Higher-order terms}}\sigma^n \tag{1.31}$$

where $\mathrm{d}^nu(\bar{R})/\mathrm{d}R^n$ is the nth derivative, σ^n is the nth moment of the distribution of returns, R, and $\bar{R}$ is the centre of this distribution. If we limit ourselves to a quadratic approximation, we see that mean–variance analysis can be regarded as a second-order Taylor approximation of expected utility.[26] Alternatively, we could solve the "true" decision problem

$$E[u(R)] = \frac{1}{T}\sum_{t=1}^{T} u(R_t) \tag{1.32}$$

where we calculate expected utility by adding up the utilities of end-of-period wealth in T equally likely scenarios, R_t (ie, the empirical return distribution).

The question now becomes: how well does Equation 1.31 approximate Equation 1.32? This depends on several factors.

Specification of the utility function. The utility of highly risk-averse investors is heavily influenced by a few extremely negative returns; hence, the approximation will perform badly for large risk aversions.

Distribution of returns. If returns are very non-normally distributed, the mean and variance no longer adequately describe the return distribution, so Equation 1.32 might assign different rankings to investment alternatives. It is also not clear that higher-order approximations do better as they only do better locally. In fact, as many returns are not infinitesimal, these approximations might even do worse. In these cases scenario optimisation might be the only viable alternative. However, as long as portfolios are diversified, we can rely on the Central Limit Theorem (see Chapter 3) to ensure that higher moments in excess of those of the normal distribution are close to zero.

Volatility. The greater the volatility of a return series (or, alternatively, the longer the time horizon), the less appropriate the Taylor approximation becomes.

In practice, the suitability of a mean–variance approximation is judged by ranking assets on the basis of Equation 1.31 and the mean–variance approximation of Equation 1.32. Then rank correlations between the two measures are calculated to see whether both place assets in the same order. Empirical results on the applicability of mean–variance-based approximation sometimes conflict as they seem to be time-period-dependent.[27] The practical applicability of mean–variance-based investing should become an empirical issue of whether the data confirm the assumptions rather than a battle between dogmatic views.

However, investors can never go wrong using direct utility optimisation. While all risk measures can be gamed (as they either ignore deviations from normality as in the case of or focus on percentiles as in the case of value at risk), utility weighs all parts of the return distribution due to their economic importance and directly accommodates for non-normality.

We could, for example, game value at risk by selling deep out-of-the-money puts. This would create a portfolio that increases in equity allocation as wealth falls, which would deeply hurt all investors with constant or decreasing relative risk aversion, but remain undetected by a value at risk measure. In utility based portfolio construction this would not remain undetected.

1.2.2 How well does the one-period solution approximate multiperiod optimality?

The second area where the facts may differ in practice relates to the ability of the mean–variance framework to deal with multiperiod (ie, long-term) investment decisions. Recall that the Markowitz framework is used by institutional investors who have very long-term horizons. We will divide this problem into three separate questions.

1. Does the mean–variance frontier change as the investment horizon lengthens?
2. Should all long-term investors maximise geometric return?
3. Does repeatedly investing in myopic (one-period-efficient) portfolios result in multiperiod-efficient portfolios?

Note that the first two questions assume that investors cannot rebalance within the investment period, so a sequence of investment periods actually becomes one very long period, while the third question allows rebalancing.

The first question is easy enough to answer. Assuming non-time-varying, uncorrelated and normally distributed returns, expected portfolio excess returns and variance are scaled up by the number of periods, T. However, this leaves the curvature of the efficient frontier (trading off return against variance) unchanged, so all investors will choose the same portfolio irrespective of the scaling (ie, time horizon), T. It is true that Sharpe ratios (average return relative to standard deviation) rise with investment horizon as standard deviation rises only as the square root of T. However, it is equally true that Sharpe ratios must be compared between assets for a given time horizon and not between time horizons for a given asset. Risk does not diversify across time.

To answer the second question, an often-quoted example (which seems to affirm the suggestion underlying the question) can be used. Suppose that an investment starts at a value of 100, falls to 50 and rises again to 100. Although the average return is 25% $((-50 + 100)/2)$, the geometric return is zero. In fact, repeatedly investing in this portfolio would leave the expected average return constant at 25%, while the geometric return would eventually become negative. For a start, we see that geometric returns also have a risk dimension. All things being equal (same average return), the higher the volatility (the less diversified a portfolio), the lower the geometric return on the portfolio will be and, hence, the lower the expected median wealth. (Equally, the same average return will produce a less diversified portfolio.) However, as the geometric mean return, g_{p}, can be approximated by $g_{\mathrm{p}} \approx \mu_{\mathrm{p}} - \frac{1}{2}\sigma_{\mathrm{p}}^2$, where μ_{p} is portfolio return, we know that the portfolio that maximises geometric mean is merely a particular frontier portfolio.[28] It is represented by the point on the efficient frontier where the risk-tolerance parameter, λ, equals one and reflects a fairly aggressive investor. More aggressive portfolios, as well as less aggressive portfolios, would result in a lower geometric return. Investors who are concerned that mean–variance-efficient portfolios might reduce the long-term geometric returns are often recommended to ignore allocations lying to the right of this point. We now know that not every investor will maximise geometric returns, but what type of investor would? Suppose we engage in an investment where we either win 100% in T_{win} out of T cases (periods) or lose 100% in the remaining cases. Suppose that our initial wealth is one. How much would we bet in each draw to maximise the geometric return (assuming that $T_{\text{win}}/T > 0.5$)? Betting everything in every period would almost surely result in a total loss, while betting too little might not let our capital grow fast enough. The geometric return for the problem described so far can be written as

$$g = (W_T)^{1/T} = (1+b)^{T_{\text{win}}/T}(1-b)^{(T-T_{\text{win}})/T} \tag{1.33}$$

where W_T is end-of-period wealth at time T and b is the bet size (how much is invested per period). Taking logarithms, we see that Equation 1.31 becomes the expected utility maximisation for a log-utility investor

$$\begin{aligned}\log(W_T)^{1/T} &= \frac{T_{\text{win}}}{T}\log(1+b) + \frac{T-T_{\text{win}}}{T}\log(1-b)\\ &= E[u(W)] \qquad (1.34)\end{aligned}$$

Maximising Equation 1.32 with respect to the optimal bet size, b^*, results in[29]

$$b^* = 2\frac{T_{\text{win}}}{T} - 1 \tag{1.35}$$

Choosing a higher bet size is called "over-betting" (as seen in more aggressive portfolios), and it would reduce geometric return and, hence, expected median wealth. If there is a 55% probability of winning in a single run we would, according to Equation 1.35, find it optimal to invest 10% per period. Rather than making geometric return maximisation a general objective, it is confined to log-utility investors.

Turning to the third question: suppose now that, in all future periods, we could readjust our optimal allocations at the start of each period. Under what circumstances would that affect today's decisions? Under fairly strict assumptions it has been shown that repeatedly investing in myopic (one-period-efficient) portfolios will also result in multiperiod-efficient portfolios if[30]

- investors have constant relative risk aversion (wealth level does not change optimal allocations) and only possess financial wealth;
- asset returns are not autocorrelated (investment opportunities are not time-varying), ie, period returns are not forecastable;[31]
- portfolio returns are not path-dependent due to intermediate cashflows (no cash infusion and/or withdrawals); and
- there are no transaction costs (which make optimal allocations path dependent as rebalancing now depends on the size of returns).

The last two requirements, however, are unrealistic as investment opportunities are time-varying and transaction costs are unavoidable. However, a fifth condition has only recently been been added to the above list:

- there is no uncertainty about estimated parameters.

If investors face parameter uncertainty in relation to mean returns, they might actually reduce the optimal holding of risky assets as the time horizon lengthens. However, if investors learn to update their beliefs as they observe further return realisations, they might start to increase their holdings again.[32]

As many of the implementation problems – such as estimation error (see next section) or the ability to specify a utility function – are likely to be worse in a multiperiod framework, and given the technicality of the subject, our analysis of multiperiod models will end here.[33]

1.2.3 Who should use portfolio optimisation?

Portfolio optimisation requires the decision maker to trade of return against a given measure of risk. To decide upon what trade-off is attractive the concept of utility has been introduced. With the help of a utility function (or its approximation via a suitable objective function) we can rank risky investments. While this is (at least conceptually) straightforward for individual investors, the application of portfolio theory for institutional investors[34] is highly questionable. This is because institutional investors do not exhibit utility. Corporations only live through their bond and equity investors, but do not have a life of their own. If they do we face a corporate governance problem. Corporations do not live! This line of thought is actually quite old and traces back to the Fisher–Hirschleifer Theorem known from any serious corporate finance textbook. It states that capital budgeting decisions can be made without recurring to the shareholders′ utility. In fact the decision maker can be entirely agnostic about their utility functions. All they need to do is to invest into those projects that have the highest net present value. This directly leads to modern capital budgeting where projects (also risky investments) are valued without the need to know shareholders′ utility. In fact it is one of the biggest achievements of modern finance to separate valuation and utility.

What does this mean for the application of portfolio theory to institutional investment problems? We cannot compare capital market investments on the basis of some risk return trade-off as the net present value of all capital market investments is by zero. Investing in the S&P 500 will still carry a positive risk premium but the required discount factor rises in parallel. We need other criteria outside the reach of portfolio theory to make meaningful decisions. A simple example will help to illustrate this. Assume a weak company (Double B rating) with an under-funded pension plan. Assume further the existence of a pension insurance system (like the PBGC in the USA). It is now in the best interest of the shareholder to invest 100% into risky assets (highly correlated with the firm′s operating assets) in order to exploit the pension put written by the pension guarantee company. Traditional mean variance optimisation could not arrive at this result. In short we need to apply contingent claims analysis (that also does not rely on utility) rather than a naive transfer of portfolio theory (developed for individual investors).

1.3 ANALYSIS OF IMPLIED RISK AND RETURN

1.3.1 Basic risk decomposition

For most investors a single risk number is not enough; they want to know the sources of risk and how well spread they are to enable them to judge whether a portfolio is diversified or not, and which assets are diversifying.[35] The starting point for the decomposition of total portfolio risk (sometimes also called the "risk budget") is the familiar formula for the standard deviation of portfolio returns[36]

$$\sigma_{\mathrm{p}} = (\boldsymbol{w}'\boldsymbol{\Omega}\boldsymbol{w})^{1/2} = \sum_i w_i^2\sigma_{ii} + \sum_i\sum_{j\neq i} w_i w_j \sigma_{ij} \tag{1.36}$$

The first question we want to address is: how would the portfolio risk change if we increased holdings in a particular asset? What we need is the "marginal contribution to risk" (MCTR), which can be calculated by taking the first derivative of Equation 1.36, giving

$$\mathrm{MCTR}_{k\times 1} = \frac{\mathrm{d}\sigma_{\mathrm{p}}}{\mathrm{d}\boldsymbol{w}} = \frac{\boldsymbol{\Omega}\boldsymbol{w}}{\sigma_{\mathrm{p}}} \tag{1.37}$$

where the ith element in this $k \times 1$ vector is given by

$$\frac{\mathrm{d}\sigma_{\mathrm{p}}}{\mathrm{d}w_i} = \frac{w_i\sigma_{ii} + \sum_{j\neq i} w_j\sigma_{ij}}{\sigma_{\mathrm{p}}} = \frac{\sigma_{i\mathrm{p}}}{\sigma_{\mathrm{p}}} = \beta_i\sigma_{\mathrm{p}} \tag{1.38}$$

The MCTR rises directly with asset weight, w_i, and asset riskiness, σ_{ii}. However, the net effect of an increase in weight can still be negative (falling risk) if the asset shows a sufficiently negative covariance, σ_{ij}, with the other assets.

Manipulation of Equation 1.38 shows that its numerator is equal to the covariance of asset i with portfolio p, which contains the asset

$$\sigma_{i\mathrm{p}} = \mathrm{cov}(r_i, r_p) = \mathrm{cov}\left(r_i, w_i r_i + \sum_{j\neq i} w_j r_j\right) = w_i\sigma_{ii} + \sum_{j\neq i} w_j\sigma_{ij} \tag{1.39}$$

Adding up the weighted MCTRs yields the volatility of the portfolio

$$\sum_i w_i \frac{\mathrm{d}\sigma_{\mathrm{p}}}{\mathrm{d}w_i} = \sum_i w_i \frac{\sigma_{i\mathrm{p}}}{\sigma_{\mathrm{p}}} = \sigma_{\mathrm{p}} \tag{1.40}$$

Dividing Equation 1.39 by σ_{p} yields

$$\sum_i \frac{w_i}{\sigma_{\mathrm{p}}} \frac{\mathrm{d}\sigma_{\mathrm{p}}}{\mathrm{d}w_i} = \sum_i w_i \frac{\sigma_{i\mathrm{p}}}{\sigma_{\mathrm{p}}^2} = \sum_i w_i\beta_i = 1 \tag{1.41}$$

which shows that the percentage contributions to risk (PCTR), which add up to 100%, are equal to the weighted betas. We can write this in matrix form as

$$\mathrm{PCTR}_{k\times 1} = \frac{\boldsymbol{W}}{\sigma_{\mathrm{p}}}\frac{\mathrm{d}\sigma_{\mathrm{p}}}{\mathrm{d}\boldsymbol{w}} \tag{1.42}$$

where $\boldsymbol{W}$ is a $k \times k$ matrix with portfolio weights on the main diagonal and zeros otherwise. An individual element of the vector of percentage risk contributions is given by

$$\mathrm{PCTR}_i = \frac{w_i}{\sigma_{\mathrm{p}}}\frac{\mathrm{d}\sigma_{\mathrm{p}}}{\mathrm{d}w_i} = w_i\beta_i \tag{1.43}$$

We can use Equations 1.36 to 1.43 to show that for the (unconstrained) minimum–variance portfolio all PCTRs have to equal the respective portfolio weights. For such a portfolio we know that all MCTRs have to be equal by definition, ie

$$\frac{\mathrm{d}\sigma_{\mathrm{p}}}{\mathrm{d}w_i} = \frac{\mathrm{d}\sigma_{\mathrm{p}}}{\mathrm{d}w_j}$$

If this were not so, we could always find a pair of assets where slightly increasing one holding while at the same time reducing the other would result in lowered risk. Using the fact that $\sum w_i = 1$, we can now show that, using Equation 1.39

$$\sum_i w_i \frac{\mathrm{d}\sigma_{\mathrm{p}}}{\mathrm{d}w_i} = \sigma_{\mathrm{p}} \quad \text{and} \quad \sum_i \frac{w_i}{\sigma_{\mathrm{p}}}\frac{\mathrm{d}\sigma_{\mathrm{p}}}{\mathrm{d}w} = w_i \tag{1.44}$$

PCTRs are also used as an indication of the level of diversification of a portfolio: for example, if 90% of the total risk comes from a few holdings, a portfolio is probably not very diversified. However, that does not reflect on the standard of the portfolio as diversification is not a value in itself.[37]

1.3.2 Implied view analysis: reverse optimisation

So far we have calculated optimal portfolio weights from given return expectations. Often, however, the only givens are portfolio weights. If investors have high return expectations for assets with high marginal risk contributions, how can we find out whether their portfolios truly reflect their views and whether what they invest in is really based on their opinions? As we all know, conviction is the natural enemy of diversification.

Table 1.1 Implied view analysis: assumptions

	Weight	Return	Volatility							
Asset	**(%)**	**(%)**	**(%)**	**Correlation**						
Equity	40	11	18	1.0	0.0	0.5	0.5	0.3	0.3	0.0
Absolute return	15	12	8	0.0	1.0	0.0	0.0	0.0	0.0	0.0
Private equity	15	11	9	0.5	0.0	1.0	0.5	0.3	0.3	0.0
Real estate	5	10	14	0.5	0.0	0.5	1.0	0.5	0.3	0.0
US bonds	25	7	3	0.3	0.0	0.3	0.5	1.0	0.8	0.0
Non-US bonds	0	8	8	0.3	0.0	0.3	0.3	0.8	1.0	0.0
Cash	0	5	0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

Column heads for correlation matrix read as items in “Asset” column, ie, from “Equity” (left) to “Cash” (right).

This can be done by “reverse optimisation” – a technique of mapping positions on to implicit return expectations. In an unconstrained portfolio optimisation, marginal risks (additional risks that arise through extending a position) are traded off against marginal returns (additional gains from extending a position). A portfolio is thus optimal when the relationship between marginal risks and marginal returns is the same for all assets in the portfolio. In this case, changing assets cannot generate a better risk–return relationship. However, as the Sharpe ratio of the portfolio also measures the relationship between incremental risk and return, we can express the relationship between marginal return and marginal risk as

$$\boldsymbol{\mu} = \left(\frac{\mu_{\mathrm{p}}}{\sigma_{\mathrm{p}}}\right)\frac{\boldsymbol{\Omega w}}{\sigma_{\mathrm{p}}} = \boldsymbol{\beta}\mu_{\mathrm{p}} \tag{1.45}$$

where $\boldsymbol{\beta} = \boldsymbol{\Omega w \sigma}_{\mathrm{p}}^{-2}$ denotes the $k \times 1$ vector of asset betas. Betas measure the sensitivity of an asset to movements of the portfolio. High-beta assets have to earn a high risk premium to justify the risk contribution to a particular portfolio. Note that Equation 1.45 follows from portfolio mathematics and not from an equilibrium condition. However, if the analysed portfolio were the market portfolio, we could give it an equilibrium interpretation. As the market portfolio is often conveniently proxied by a capitalisation-weighted index of all publicly tracked assets, the implied returns would be the returns investors would need to hold the market portfolio. As capital market theory predicts that all investors hold the market portfolio, we could interpret these as equilibrium returns. This kind of analysis can be

Table 1.2 Implied view analysis: risk contributions and implied returns

Asset	PCTR$_i$ (%)	MCTR$_i$	Implied return (%)
Equity	79.1	0.174	9.84
Absolute return	1.9	0.011	0.62
Private equity	10.2	0.060	3.39
Real estate	4.8	0.085	4.80
US bonds	4.0	0.014	0.80
Non-US bonds	0.0	0.029	1.66
Cash	0.0	0.000	0.00

PCTR$_i$, percentage contribution to risk of ith asset; MCTR$_i$, marginal contribution to risk of ith asset. Implied return given by Equation 1.45.

employed to show investors whether their return expectations are consistent with market realities, ie, whether they are over- or underspending their risk budget in particular areas and whether they are investing in a way that is consistent with their views.

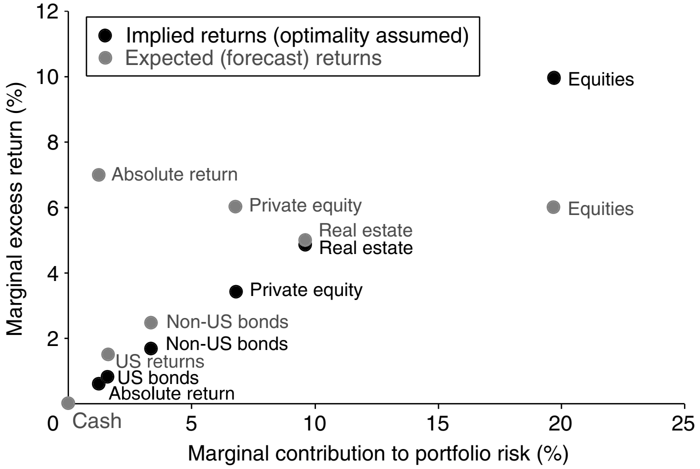
We will illustrate the reverse optimisation procedure with an example. Suppose that a US-dollar-based investor has the views on the world set out in Table 1.1. The annual risk on this portfolio accumulates to 8.79% with an expected return of 10% (excess return of 5%) and a Sharpe ratio of 0.57. Risk contributions according to Equations 1.38 and 1.42 are given in Table 1.2.

What does a marginal risk contribution of 0.014 for US bonds mean? Suppose that instead of holding 25% (as in Table 1.1), we invested 26% in US bonds (we can assume that this extra 1% has been funded from cash; see also Chapter 8 on funding assumptions).[38] Total portfolio risk would change from 8.7948 to 8.8089, a difference of 0.0141

$$\begin{aligned}\Delta\sigma_{\mathrm{p}} &= 8.8089 - 8.7948 = 0.0141\\ &= \frac{\mathrm{d}\sigma_{\mathrm{p}}}{\mathrm{d}w_{\text{US bonds}}}\Delta w_{\text{US bonds}}\end{aligned}$$

The biggest increase in risk would come from equities (which already account for about 80% of the total risk budget), while the smallest increase would come from absolute return strategies, which are the most diversifying asset (under the set of assumptions given in Table 1.1). With these calculations in mind, we are now well equipped to analyse the implied views using Equation 1.45.

Figure 1.2 Implied view analysis – comparison of implied and expected returns



We see in Figure 1.2 that implied returns lie on a straight line with a slope equal to the portfolio Sharpe ratio. Return forecasts and actual allocations (risk allocations) in the figure are not consistent (if they were perfectly consistent they would overlap); the implied excess return for absolute return strategies (0.6%) is lower than forecast (7%). This means that the investor is underspending risk in this area. The opposite is true for equities, where the investor overspends risk (relative to the optimal allocation) for this asset class. A large allocation in a relatively undiversifying asset requires large implied returns to make the portfolio optimal. It also becomes apparent that the investor's implied return for equities (return required to make the position worthwhile) is higher than the historical experience.

Implied returns are often employed in what is called "view optimisation". This is not an optimisation but rather an iterative process of trial and error where an investor changes allocations (and sometimes forecasts) until implied and expected returns show a "reasonable" overlap. Implied returns could also be used to create a global firm view by optimally combining the views of regional model portfolio teams. However, this should be used with caution as implied views only show the view of an unconstrained investor. For example, if an investor is not allowed to hold more than 10% of a particular equity market, the implied views are also capped. This issue also applies to long-only constraints. Negative views on an asset with 2% index weight are capped as the maximum short position is 2% (ie, not owning this asset at all). If these issues are not addressed, mixing views and constraints contaminates the global firm view.

1.3.3 Implied views with liabilities

How will the presence of liabilities change our previous analysis? Using our previous notation we can express the "utility" (the author strongly objects to this kind of pension fund decision making, but includes it for completeness) for a stand-alone pension fund as

$$\max_{\boldsymbol{w}} u = \boldsymbol{w}' - \frac{1}{2\lambda}(\boldsymbol{w}'\boldsymbol{\Omega}\boldsymbol{w} - 2f\boldsymbol{w}'\boldsymbol{\Gamma}) \tag{1.46}$$

Taking the first derivative of Equation 1.46 with respect to $\boldsymbol{w}$ and solving for $\boldsymbol{\mu}$ yields

$$\boldsymbol{\mu}_{\text{impl}} = \lambda^{-1}(\boldsymbol{\Omega}\boldsymbol{w} - f\boldsymbol{\Gamma}) = \begin{pmatrix}\text{mean variance}\\ \text{implied returns}\end{pmatrix} - \frac{f}{\lambda}\boldsymbol{\Gamma} \tag{1.47}$$

We see that the usual solution for implied returns is extended by a liability hedging credit. This term has three influence factors.

- Assets that have high covariance ($\boldsymbol{\Gamma}$) with liabilities need to return less than assets with less favourable hedging properties.
- If risk aversions is high (small λ), liability hedging will become even more important, in that assets that have high covariance with liabilities require even less return.
- For pension plans with large surpluses (small f), liability hedging becomes less important and so the importance of the liability hedging is reduced.

Conventional thinking is reversed. Assets with large correlation (with liabilities) require a small risk premium. This is different from traditional analysis, where large correlations demand high-risk premiums. Remember that low correlations with liabilities increases the likelihood that assets and liabilities drift apart, ie, they increase surplus risk rather than reducing it.

1.3.4 Applying investment information consistently across client portfolios

A final application of implied views is the so-called "portfolio-factory".[39] Typically, an asset manager has nearly as many benchmarks as clients. To complicate matters further, clients' guidelines differ. How can the asset manager ensure that all client portfolios are treated the same way, ie, that they are all based on the same investment information? When the global model portfolio has been specified, how can portfolio construction for each client be automated? One way (and probably the only way to achieve this in its purest meaning) is the following three-step procedure.

Step 1. Construct a totally unconstrained model portfolio (otherwise views are contaminated by constraints).

Step 2. Back out the implied returns (solve for the returns that make current weights optimal).

Step 3. Run a standardised portfolio optimisation with client-specific constraints for each client whereby the same implied returns are used for all clients.

Following this process, client portfolios would still differ in (active and total) weights but not in input information.

1.4 RISK BUDGETING VERSUS PORTFOLIO OPTIMISATION

1.4.1 The problem

Risk budgeting is an increasingly fashionable topic. The reason for this is clear: for a pension fund manager not to engage in risk budgeting would seem almost negligent. Also, as institutional investors are disappointed with the economic value that traditional portfolio optimisation methods have provided, they are more willing to embrace a budgeting framework that allows them to plan and spend risk budgets.

Many authors associate risk budgeting with value-at-risk (VaR) as a measure of investment risk.[40] In fact, supporters of risk budgeting transform monetary allocations into VaR (or marginal VaR). No additional economic insight is gained by this apart from the educational insight that even small monetary (weight) allocations can make a large contribution to total risk.

Promoters of risk budgeting would like to "separate risk budgeting and VaR measurement from classic investment risk practices, such as asset allocation",[41] while others argue that "we should regard risk budgeting as an extension of mean–variance optimisation that enables us to decouple a portfolio's allocation from fixed monetary values"[42] and, hence, that "VaR and the risk capital budgeting metaphor pour old wine into new casks".[43] With regard to the last two comments, it can additionally be argued that the merit of risk budgeting does not come from any increase in intellectual insight but rather from the more accessible way it provides of decomposing and presenting investment risks. That said, for the purposes of this book we will define risk budgeting as a process that reviews any assumption that is critical for the successful meeting of prespecified investment targets and thereby decides on the trade-off between the risks and returns associated with investment decisions. In a mean–variance world this defaults to Markowitz portfolio optimisation, where results are not only shown in terms of weights and monetary allocations but also in terms of risk contributions. Hence, as argued in the Introduction, risk budgeting should not be seen as a way of enforcing diversification. An isolated focus on risk is a meaningless exercise without trading off risk against return.

1.4.2 Equivalence of VaR-based risk budgeting

The equivalence of VaR-based measures and classic risk measures (as they have been known in portfolio theory for the last 40 years) is best shown by relating VaR to the analysis in Section 1.3. To do this we need only to transform VaR from its monetary form into a return VaR, R_{p}^{*}[44]

$$\mathrm{VaR} = \Delta P = P R_{\mathrm{p}}^{*} = P(\mu_{\mathrm{p}} + z_{\alpha}\sigma_{\mathrm{p}}) \tag{1.48}$$

where ΔP denotes changes in portfolio value P and z_{α} is the critical value from a standard normal distribution. As in the previous section, we want to know marginal VaR, ie, we ask how VaR changes if we increase the holdings in the ith asset. Again, we take the first derivative of return VaR to arrive at marginal VaR

$$\frac{\mathrm{d}R_{\mathrm{p}}^{*}}{\mathrm{d}w_i} = z_{\alpha}\frac{\mathrm{d}\sigma_{\mathrm{p}}}{\mathrm{d}w_i} = z_{\alpha}\beta_i\sigma_{\mathrm{p}} = \beta_i R_{\mathrm{p}}^{*} \tag{1.49}$$

As can be seen, the difference between the marginal contribution and risk in Equation 1.38 is merely the multiple z_{α}. The product of marginal VaR and portfolio VaR is called "component VaR" because the sum of the individual components adds up to total VaR

$$\sum_i w_i \frac{\mathrm{d}R_{\mathrm{p}}^{*}}{\mathrm{d}w_i} = \sum_i w_i\beta_i R_{\mathrm{p}}^{*} = R_{\mathrm{p}}^{*}$$

Again, this is just a multiple of Equation 1.40. Hence, in a normally distributed world, risk budgets are best derived by running a mean–variance optimisation first and subsequently transforming monetary allocations into VaR exposures.

1.4.3 Pitfalls: risk is not additive

The reverse approach, where investors first choose the VaR exposures of their activities (without trading off marginal risk and return) and then aggregate them to enforce diversification, suffers from various problems.

1. A bottom-up approach is not feasible as VaR is not additive (only being so in the case of perfectly correlated VaRs). The sum of individual VaRs will always be greater than portfolio VaR. It will therefore take many trials before a satisfactory answer is found to questions on how to spend the total risk (VaR) budget.

2. Without trading off marginal risks against marginal returns, the resulting solution is highly likely to be inefficient, ie, it carries too much risk per unit of return.

3. Individual VaR does not show the impact of a particular position on portfolio risk as correlations remain unaccounted for. Trial and error will not find trades that reduce portfolio risk in the best possible way as portfolio risk is best reduced by reducing the position of assets that have the highest marginal VaR rather than those with the highest individual VaR.

4. Focusing on VaR as a risk measure might be the best option for some, but few would claim to be indifferent when faced with a choice between a loss that matches VaR and a loss that is 10 times VaR, which is the logic underlying VaR optimisation.

However, what about non-normality? VaR-based measures start to become preferable to mean–variance-based allocations if returns are significantly non-normal or if positions involve non-linearities. However, this does not mean that risk budgeting is different from portfolio optimisation. Risk budgeting still needs to find the optimal trade-off between risk and return (later chapters show how non-normal returns can be incorporated into portfolio construction). Remarkably, proponents of risk budgeting describe in great detail how to calculate VaR and why this might be a good measure of risk, but they do not explain how to arrive at a risk budget.[45]

1.5 THE UNCONDITIONAL COVARIANCE MATRIX AND ITS PROPERTIES

1.5.1 Introduction

The covariance matrix is a fundamental tool for risk estimation and one that investment professionals routinely use to calculate portfolio variance.[46] But what condition do we have to place on the covariance matrix to be sure that $\boldsymbol{w}'\boldsymbol{\Omega}\boldsymbol{w} \geqslant \mathbf{0}$ is always true? After all, variance cannot be negative.

Fortunately we can use some well-known results from matrix algebra.[47] A matrix, $\boldsymbol{\Omega}$, that satisfies $\boldsymbol{w}'\boldsymbol{\Omega}\boldsymbol{w} \geqslant \mathbf{0}$ for all $\boldsymbol{w}$ is referred to as "positive semi-definite". A necessary and sufficient condition for positive semi-definiteness (for symmetric matrixes) is that all the eigenvalues of $\boldsymbol{\Omega}$ are positive or zero and at least one eigenvalue is greater than zero. This can be checked by obtaining the solutions to the eigenvalue equation $\boldsymbol{\Omega}\boldsymbol{x} = e\boldsymbol{x}$, where $\boldsymbol{x}$ is a $k \times 1$ vector ("eigenvector") and e is a scalar ("eigenvalue"). There are k solutions, $e_1, \dots, e_k$ (eigenvalues), to this equation. If all the solutions (roots) are positive or zero, and at least one is positive, the matrix is said to be positive semi-definite.

Consider the following simple covariance matrix (to facilitate interpretation we will assume unit variances)

$$\boldsymbol{\Omega} = \begin{bmatrix} 1 & 0.9 & 0.8 \\ 0.9 & 1 & 0.7 \\ 0.8 & 0.7 & 1 \end{bmatrix}$$

We calculate the following eigenvalues[48]

$$(e_1, e_2, e_3) = (2.6, 0.31, 0.08)$$

As none of them is negative, the above covariance matrix is positive semi-definite. However, if we construct an obviously contradictory case

$$\boldsymbol{\Omega} = \begin{bmatrix} 1 & 0.9 & -0.3 \\ 0.9 & 1 & 0.7 \\ -0.3 & 0.7 & 1 \end{bmatrix}$$

where variables one and two as well as two and three are highly positively correlated but one and three are negatively correlated, we find that one of the eigenvalues, $(e_1, e_2, e_3) = (2, 1.28, -0.3)$, is negative. This situation typically arises if we fabricate numbers, if we generate estimates from time series of different lengths or if the number of observations is smaller than the number of assets (risk factors).[49] So, what can we do to correct this problem with minimum damage to the input structure? The three-step process below is one possible method.[50]

1. Find the smallest eigenvalue (here e_3).
2. Create the minimum eigenvalue of zero by shifting the correlation matrix according to $\boldsymbol{\Omega}^* = \boldsymbol{\Omega} - e_3\boldsymbol{I}$, where $\boldsymbol{I}$ denotes the identity matrix.
3. Scale the resulting matrix by $1/(1 - e_3)$ to enforce for each variable a correlation with itself of one

$$\boldsymbol{\Omega}^{**} = \frac{1}{1 - e_3}\boldsymbol{\Omega}^*$$

The result is the new adjusted matrix

$$\boldsymbol{\Omega}^{**} = \begin{bmatrix} 1 & 0.69 & -0.23 \\ 0.69 & 1 & 0.54 \\ -0.23 & 0.54 & 1 \end{bmatrix}$$

with eigenvalues $(e_1^{**}, e_2^{**}, e_3^{**}) = (1.77, 1.22, 0)$. Alternatively, we can obtain the same results as in the example above if we use the fact that $\boldsymbol{\Omega} = \boldsymbol{XEX}'$, where $\boldsymbol{X}$ is the $k \times k$ matrix of eigenvectors and $\boldsymbol{E}$ is a $k \times k$ diagonal matrix (eigenvalues on the main diagonal and all other terms being zero) of eigenvalues. If we set all negative eigenvalues to zero $(\boldsymbol{E}^*)$, we can calculate $\boldsymbol{\Omega}^* = \boldsymbol{XE}^*\boldsymbol{X}'$. To ensure correlation values of one on the main diagonal, we calculate

$$\boldsymbol{\Omega}^{**} = \frac{1}{\sqrt{\boldsymbol{D}}}\frac{\boldsymbol{\Omega}^{*1}}{\sqrt{\boldsymbol{D}}}$$

where $\boldsymbol{D}$ is a diagonal matrix containing the main diagonal of $\boldsymbol{\Omega}^*$.

1.5.2 Random matrix theory

Practitioners have long been concerned about noisy estimates of the variance covariance matrix. For example, looking at the covariance matrix of active manager returns (where common market factors have been more or less removed by subtracting benchmark returns) the off-diagonal elements of the covariance matrix suspiciously look close to zero, even though empirically they are not. However, using these non-zero inputs in portfolio optimisation might yield to biased risk forecasts as the optimiser adjusts to noise rather than information. Hence, two questions naturally arise for the practitioner.

- How can we judge whether a given correlation matrix (standardised covariance matrix) is statistically different from a correlation matrix that exhibits zero correlation (also called random correlation matrix)?
- How can we clean a given correlation matrix to separate the signal from the noise, ie, how can we identify the true information (signal) in a noisy correlation matrix?

Random matrix theory (RMT) tries to answer these questions.[51] Under the restriction that $T, k \to \infty T$ and $q = T/k > 1$ it has been shown[52] that the distribution of eigenvalues for a random correlation matrix is given by

$$p(e) = \begin{cases} \dfrac{q}{2\sigma\pi} \dfrac{\sqrt{(e_{\max} - e)(e - e_{\min})}}{e} & e_{\max} \geqslant e \geqslant e_{\min} \\ 0 & \text{else} \end{cases} \tag{1.50}$$

where

$$e_{\max,\min} = \sigma^2\left(1 + \frac{1}{q} \pm 2\sqrt{q^{-1}}\right) \tag{1.51}$$

and σ^2 is the variance of the elements of the $T \times k$ data matrix. In the case of using normalised time series (to directly arrive at the correlation matrix) $\sigma^2 = 1$ by definition. In other words, if the eigenvalues of the empirical correlation matrix differ from that bound we can say that they do not come from a random correlation matrix and as such contain information. Collecting 2,000 data points (38.46 years of weekly data) for a universe of 1,500 assets (MSCI World), leads to $e_{\max,\min} = [3.482, 0.018]$ for the empirical correlation matrix.

Now that we can select the meaningful eigenvalues, how can we clean the correlation matrix, ie, how can we use RMT to remove the noise? A simple procedure is the following.

1. Calculate the empirical correlation matrix together with its (ordered) eigenvalues.
2. Compare the largest eigenvalue with the boundary in Equation 1.51.[53]
3. Count the number of significant eigenvalues.

The reader will recognise that RMT essentially describes an algorithm how to determine the number of principal components, cutting off those principal components that do not contain information that can be distinguished from noise.

While RMT offers (despite its mysterious name) a straightforward mechanism for correlation cleansing, it does not always deliver desired results. A variety of authors have shown that eigenvalues in the bulk of the spectrum (between $e_{\max}$ and $e_{\min}$) can contain significant information.[54] Moreover, there is a tendency to underestimate the largest correlations[55], which is particularly troubling for the risk management of long-only mandates.

1.5.3 Significance of inverse of the covariance matrix for portfolio construction

Now we have discussed conditions for the existence of a well-specified covariance matrix, we can throw further light on the economic interpretation of the inverse of the covariance matrix, $\boldsymbol{\Omega}^{-1}$, with regard to portfolio construction. From Section 1.1 we know the solution to the (unconstrained) portfolio optimisation problem $\boldsymbol{w} = \lambda\boldsymbol{\Omega}^{-1}\boldsymbol{\mu}$. It has been shown that, after tedious manipulation, the inverse of the covariance matrix can also be expressed as[56]

$$\boldsymbol{\Omega}^{-1} = \begin{bmatrix} \dfrac{1}{\sigma_{11}(1-R_1^2)} & -\dfrac{\beta_{12}}{\sigma_{11}(1-R_1^2)} & \cdots & -\dfrac{\beta_{1k}}{\sigma_{11}(1-R_1^2)} \\ -\dfrac{\beta_{21}}{\sigma_{22}(1-R_2^2)} & \dfrac{1}{\sigma_{22}(1-R_2^2)} & \cdots & -\dfrac{\beta_{2k}}{\sigma_{22}(1-R_2^2)} \\ \vdots & \vdots & \ddots & \vdots \\ -\dfrac{\beta_{k1}}{\sigma_{kk}(1-R_k^2)} & -\dfrac{\beta_{k2}}{\sigma_{kk}(1-R_k^2)} & \cdots & \dfrac{1}{\sigma_{kk}(1-R_k^2)} \end{bmatrix} \tag{1.52}$$

Going through the above notation, we again write the variance of the ith asset as σ_{ii}, while the betas of a return regression of asset i against all other $k-1$ assets are denoted as β_{ij}

$$r_i = a + \sum_{j \neq i} \beta_{ij} r_j + \varepsilon_i \tag{1.53}$$

The explanatory power of this regression is given as R_i^2. Inserting Equation 1.43 into the solution for optimal portfolio weights and expanding the terms, we obtain the optimal weight for the ith asset

$$w_i^* = \lambda\left(\frac{\overbrace{\mu_i - \sum_j \beta_{ij}\mu_j}^{\text{Excess return after regression hedging}}}{\underbrace{\sigma_{ii}(1 - R_i^2)}_{\text{Non-hedgable risk}}}\right) \tag{1.54}$$

The numerator in Equation 1.54 shows the excess return after regression hedging – that is, the excess return after the reward for implicit exposure to other assets has been taken out – and is equivalent to a in Equation 1.53. As the total risk of an asset is given by σ_{ij}, the fraction of risk that cannot be hedged by other assets is $\sigma_{ii}(1 - R_i^2)$. In terms of Equation 1.53 this is the unexplained variance, which is the variance of the error term. As the regression in Equation 1.53 tries to minimise the variance of the error term, Equation 1.54 will put maximum weight into those assets that are similar to all other assets but have a very small return advantage. It is this property that will lead to implausible results when estimation errors are taken into account (see Chapter 4).

1.6 CASE STUDY: INTRODUCING OUTSIDE WEALTH

We view the optimal asset allocation problem of a sovereign wealth fund (SWF) as the decision-making problem of an investor with nontradable endowed wealth (oil reserves). In order to get insight into the portfolio choice problem for an SWF, we assume the following analytical setup.

The SWF can invest its financial wealth into a single asset or cash. We can think of this as the choice between the global market portfolio and cash. This is certainly restrictive, but will allow us to develop our framework without the need for very complex calculations and we will relax this assumption in the following section. Returns for this performance asset are normally distributed and given by

$$\tilde{r}_{\text{a}} \sim N(\mu_{\text{a}}, \sigma_{\text{a}}^2) \tag{1.55}$$

where μ_{a} represents the expected risk premium (over local cash returns) of our performance asset and σ_{a} its volatility. At the same time the government budget moves with changes on its claim on economic net wealth. For a commodity- (oil-) based SWF, changes in commodity (oil) prices will by far have the biggest influence on the government budget measured in economic (not accounting) terms. We assume that oil price changes are also normally distributed

$$\tilde{r}_{\mathrm{o}} \sim N(\mu_{\mathrm{o}}, \sigma_{\mathrm{o}}^2) \tag{1.56}$$

and correlate positively with asset returns, ie, $\operatorname{cov}(\tilde{r}_{\mathrm{a}}, \tilde{r}_{\mathrm{o}}) = \rho_{\mathrm{a,o}} > 0$. As μ_{o} is empirically extremely noisy to estimate, we look for an economic prior. Given that under perfectly integrated capital markets the Hotelling–Solow rule states that natural resource prices should grow at the world interest rate such that countries are indifferent between depletion (earning the interest rate) and keeping oil underground (earning price changes). We hence assume a risk premium on oil of zero, ie, $\mu_{\mathrm{o}} = 0$. Brent prices on January 4, 1982, were 35.9 and rose to 66.6 on October 21, 2008. This amounts to a meagre 2.3% return per annum over 26.8 years, which makes our assumption of a zero risk premium on underground oil suddenly look much more realistic. Even if we instead used the maximum oil price of 145.61, this would still amount to a mere 5.3% return, which is even more in line with average money market returns.

How do we integrate oil wealth into a country's budget surplus (deficit)? Let θ denote the fraction of importance the SWF plays in the economies government budget. A simple way to gauge this is the following consideration. If the SWF has a size of 1 monetary unit, while the market value of oil reserves amounts to 5 monetary units, this translates into $\theta = \frac{1}{1+5} = \frac{1}{6}$ weight for the SWF asset and $1 - \theta = 1 - \frac{1}{6} = \frac{5}{6}$ weight for oil revenues.[57] In other words

$$\tilde{r} = \theta w \tilde{r}_{\mathrm{a}} + (1 - \theta)\tilde{r}_{\mathrm{o}} \tag{1.57}$$

Note that $1-w$ represents the implied cash holding that carries a zero risk premium and no risk in a one-period consideration. Expressing returns as risk premium has the advantage that we do not need to model cash holdings. These simply become the residual asset that ensures portfolio weights add up to one without changing risk or (excess) return.

Suppose now the SWF manager is charged to maximise the utility of total government wealth rather than narrowly maximising the utility for its direct assets under management. The optimal solution for this problem can be found from

$$\max_{w}(\theta w \mu_{\mathrm{a}} - \tfrac{1}{2}\lambda[\theta^2 w^2 \sigma_{\mathrm{a}}^2 + (1-\theta)^2\sigma_{\mathrm{o}}^2 + 2w\theta(1-\theta)\rho\sigma_{\mathrm{a}}\sigma_{\mathrm{o}}]) \tag{1.58}$$

Taking first-order conditions and solving for w we arrive at the optimal asset allocation for a resource based SWF

$$w^* = w_{\rm s}^* + w_{\rm h}^* = \frac{1}{\theta}\frac{\mu_{\rm a}}{\lambda\sigma_{\rm a}^2} - \frac{1-\theta}{\theta}\frac{\rho\sigma_{\rm o}}{\sigma_{\rm a}} \tag{1.59}$$

Total demand for risky assets can be decomposed into speculative demand $w_{\rm s}^*$ and hedging demand $w_{\rm h}^*$. In the case of uncorrelated assets and oil resources, the optimal solution is equivalent to a leveraged (with factor $1/\theta$) position in the asset-only maximum Sharpe ratio portfolio or in other words $w_{\rm s}^*$. What is the economic intuition for this leverage? For $\theta = \frac{1}{6}$ investors with constant relative risk aversion the optimal weight of risky assets will be independent from their wealth level, which nowhere enters Equation 1.58. While a given country might have little in financial wealth in the form of SWF financial assets, it might be rich in natural resources and as such it requires a large multiplier. For we would require the SWF to leverage substantially (six times). Assuming $\mu_{\rm a} = 5$, $\sigma_{\rm a} = 20$, $\lambda = 0.03$ we get

$$\frac{1}{\theta}\frac{6}{0.03 \times 20^2} = 250\%$$

The second component in Equation 1.59 represents hedging demand. In other words, the desirability of the risky asset does not only depend on the Sharpe ratio but also on its ability to hedge out unanticipated shocks to oil wealth. Hedging demand is given as the product of leverage and oils asset beta, $\beta_{\rm o,a} = \rho\sigma_{\rm o}/\sigma_{\rm a}$. The latter is equivalent to the slope coefficient of a regression of (demeaned) asset returns against (demeaned) oil returns, ie, of the form

$$(r_{\rm o} - \bar{r}_{\rm o}) = \beta_{\rm o,a}(r_{\rm a} - \bar{r}_{\rm a}) + \varepsilon \tag{1.60}$$

Hedge demand is only zero if oil price risk is purely idiosyncratic. For $\sigma_{\rm o} = 40$, $\rho = 0.1$, we would reduce the allocation in the risky asset according to

$$-\frac{1-\frac{1}{6}}{\frac{1}{6}}\frac{0.1 \times 40}{20} = -100\%$$

Positive correlation between asset and oil price risk increases the volatility of total wealth. A 100% short position in the risky asset helps to manage total risk. However, in the case of negative correlation we would even further increase the allocation to the risky asset. The optimal position of the SWF would be 1.5 times leverage in the global market portfolio.

While the focus of this section is not on empirical work, we should provide some indication on the oil shock hedging properties of traditional asset classes. Without the existence of these assets that could potentially help to reduce total wealth volatility for oil rich investors, Equation 1.59 would be of little practical use. Let us look at global equities and (MSCI World in US dollars) and US government bonds (Lehman US Treasury total return index for varying maturities) and oil (Crude Oil–Brent Cur Month FOB from Thompson) for the period from January 1997 to September 2008. The selection of the above-mentioned assets is motivated by some basic economic considerations. Oil tends to do well either in a political crisis (in which equities do not do well) or in anticipation of global growth (in which equities also do well). At the same time government bonds (particularly at the long end) are a natural recession hedge and will also do well if oil prices fall. There are obviously notable exceptions. Oil and bonds will move together if an oil price increase is the cause of recession fears. In this scenario shorter bonds should provide better returns than long bonds due to rising inflation fears. All asset classes move together if monetary loosening inflates a leverage-driven bubble that drives equity and bond markets while bonds perform due to falling interest rates.

However, we could talk endlessly about our economic priors, so we should have a look at the data instead. The results of our correlation analysis are given in Table 1.3. In the short run (monthly data) we do not find significant correlations between oil price change and the selected asset class return. However, reducing the data frequency, ie, increasing the period to calculate returns from shows significantly negative correlations between oil price changes and fixed income returns. In other words, we find that long-term correlations are buried under short-term noise.

Both the degree of (negative) correlation and its significance (even though we reduce the sample size) rise as we move from quarterly to annual. Global equities, however, provide no hedge against oil price changes. While they could still be used as a performance asset, they are of limited use as a hedge against oil price shocks.

Proponents of equity investments might suspect that we are underselling their case as we have not been allowing for more granular equity exposures. Maybe we can identify various sectors that respond differently to oil price shocks. A global equity portfolio is

Table 1.3 Correlation of asset returns with percentage oil price changes

	US Treasury Bond							
Frequency	**1 year**	**1–3 years**	**3–5 years**	**5–7 years**	**7–10 years**	**10–20 years**	**20+ years**	**Global equities**
Monthly	−8.28%	−2.38%	1.16%	2.26%	2.75%	1.18%	0.16%	9.19%
	−0.99	−0.43	0.21	0.41	0.49	0.21	0.03	1.66
Quarterly	−9.29%	**−22.59%**	**−20.08%**	**−21.00%**	**−19.60%**	**−23.92%**	**−24.42%**	**−9.34%**
	−0.65	−2.42	−2.14	−2.24	−2.09	−2.57	−2.63	−0.98
Annual	−38.81%	**−40.70%**	**−50.94%**	**−56.52%**	**−59.48%**	**−55.04%**	**−51.29%**	**24.10%**
	−1.52	−2.36	−3.13	−3.63	−3.92	−3.49	−3.16	1.31

The table uses global equities (MSCI World in US dollars) and US government bonds (Lehman US Treasury total return index for varying maturities) and oil (Crude Oil–Brent Cur Month FOB from Thompson) for the period from January 1997 to September 2008. This translates into 142 monthly, 48 quarterly and 13 annual data points. For each data frequency, the first line shows the correlation coefficient, while the second line provides its t value. We calculate t values according to

$$t = \rho\sqrt{(n-2)/(1-\rho^2)}$$

where n represents the number of data points and ρ the estimated correlation coefficient. Critical values are given by the t-distribution with $n-2$ degrees of freedom. For example, the critical value for 13 annual data points at the 95% level is 2.2. All significant correlation coefficients are in bold.

Table 1.4 Correlation of US industry returns with percentage oil price changes

	Dow Jones industries									
Frequency	Oil & Gas	Basic materials	Industrials	Consumer goods	Health care	Consumer services	Telecom	Utilities	Financials	Technology
Monthly	**45.68%**	5.94%	−**12.89%**	−**18.03%**	−**13.83%**	−**21.26%**	−1.62%	−6.32%	−**18.08%**	−**11.86%**
	9.23	1.07	−2.34	−3.29	−2.51	−3.91	−0.29	−1.14	−3.30	−2.15
Quarterly	49.57%	8.11%	−6.47%	−**4.17%**	−**28.57%**	−22.69%	−4.55%	−6.20%	−**24.99%**	−**25.03%**
	5.96	0.85	−0.68	−0.44	−3.11	−2.43	−0.48	−0.65	−2.69	−2.70
Annual	51.03%	−20.03%	−11.99%	−**65.36%**	−**37.26%**	−**51.40%**	−**44.18%**	−15.84%	−13.52%	−3.47%
	3.14	−1.08	−0.64	−4.57	−2.12	−3.17	−2.61	−0.85	−0.72	−0.18

The table uses Dow Jones sector returns and oil (Crude Oil–Brent Cur Month FOB from Thompson) for the period from January 1982 to September 2008. This translates into 323 monthly, 109 quarterly and 28 annual data points. For each data frequency, the first line shows the correlation coefficient, while the second line provides its t value. We calculate t values according to

$$t = \rho\sqrt{(n-2)/(1-\rho^2)}$$

where n represents the number of data points and ρ the estimated correlation coefficient. Critical values are given by the t distribution with $n - 2$ degrees of freedom. For example, the critical value for 28 annual data points at the 95% level is 2.05. All significant correlation coefficients are in bold.

already a diversified portfolio that leaves no possibility of leveraging these effects. The results for this can be found in Table 1.4. Our results are encouraging. We find significant negative correlation for defensive consumer and health care sectors that tend to do well when the economy does badly. Results are stable and significant for different data frequencies. At the same time the energy sector is positively related to oil and does not qualify for inclusion into SWF allocations as we would have conjectured before.

APPENDIX A: RELATIVE MAGNITUDE OF ESTIMATION ERRORS

We have already established that estimation error on mean returns is a critical factor in portfolio construction.[58] However, as estimation error applies equally to means and variances, we want to establish the relative magnitudes of the estimation errors calculated for them.

Suppose we have T years of data. This period is divided into n intervals of equal length ($n = T/\Delta t$). What can we do to reduce the estimation error for the mean return and the uncertainty about deviations from that mean? We can either try to extend the data period (ie, increase T), or we can concentrate on shorter frequencies (ie, increase n by reducing Δt). Although the first method is often neither feasible (some series might not have a longer history) nor recommendable (a reduction in sampling error often comes at the expense of an increase in non-stationarity, as discussed in Chapter 4), it is still a useful exercise to show how much data is needed to arrive at reasonably small estimation errors.

Suppose that a fairly typical calculation gives an estimated mean of 10% and a return volatility around that mean of about 20%. Tables 1.5 and 1.6 give the results of the relevant calculations. We can see from the first table that even with 50 years of data the confidence interval on the mean estimate (11%) is the same size as the mean estimate itself.[59]

It is apparent from Table 1.6 that the estimation error on the variance of returns is many times smaller than that on mean returns.[60] However, although the estimation error becomes smaller as we increase the sample frequency, this is not necessarily the best route to pursue as increasing the frequency leads to greater autocorrelation. Moreover, daily data will exhibit an artificially low correlation between assets of different time zones as only a small number

Table 1.5 Effect of sample period on estimation error for mean returns

Estimation period, *T* (years)	Estimation error (%)	95% confidence interval (%)
1	20	78
5	9	35
10	6	25
20	4	18
50	3	11

Table 1.6 Effect of sample period on estimation error for variance estimates

Estimation period, *T* (years)	Estimation frequency			
	Daily	Weekly	Monthly	Quarterly
1	0.35	0.79	1.71	3.27
5	0.16	0.35	0.74	1.30
10	0.11	0.25	0.52	0.91
20	0.08	0.18	0.37	0.64
50	0.05	0.11	0.23	0.40

Tabulated data is estimation errors (%).

of transactions contain the same information. However, the much smaller estimation error on variance relative to that on means is the justification for the standard practice of calculating risk estimates from historical figures.

Estimation errors on means and covariances often interact in a way that is difficult to understand. Estimation error in portfolio weights is therefore difficult to attribute to either or both. Although it is tempting to declare estimation error in means as uniquely important for all investors, it is worth noting that the importance of risk estimates increases with an investor's risk aversion. This is equally true for passive investors who have to rely on risk estimates to find the best liability-mimicking portfolio. Chapter 4 deals with an interesting heuristic for visualising the uncertainty in portfolio weights created by the uncertainty in inputs.

APPENDIX B: INTRODUCING CONSTRAINTS

To be more realistic, we introduce general (binding) linear constraints of the form $\boldsymbol{A}\boldsymbol{w} = \boldsymbol{b}$, where $\boldsymbol{A}$ denotes a matrix with k columns (equal to the number of assets) and m rows (equal to the number of equality constraints) and where $\boldsymbol{b}$ is a $k \times 1$ vector of limits.[61] We maximise

$$\text{Utility} \approx \boldsymbol{w}'\boldsymbol{\mu} - \frac{1}{2\lambda}\boldsymbol{w}'\boldsymbol{\Omega}\boldsymbol{w} \quad \text{subject to } \boldsymbol{A}\boldsymbol{w} = \boldsymbol{b} \tag{1.61}$$

Forming the standard Lagrangian

$$L = \boldsymbol{w}'\boldsymbol{\mu} - \frac{1}{2\lambda}\boldsymbol{w}'\boldsymbol{\Omega}\boldsymbol{w} - \boldsymbol{\gamma}'(\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b})$$

where γ is the $m \times 1$ vector of Lagrangian multipliers (one for each constraint), and taking the first derivatives with respect to the optimal weight vector and the vector of multipliers yields

$$\frac{\mathrm{d}L}{\mathrm{d}\boldsymbol{w}} = \boldsymbol{\mu} - \frac{1}{\lambda}\boldsymbol{\Omega}\boldsymbol{w} - \boldsymbol{\gamma}'\boldsymbol{A} = \boldsymbol{0}, \qquad \boldsymbol{w}^* = \lambda\boldsymbol{\Omega}^{-1}(\boldsymbol{\mu} - \boldsymbol{\gamma}'\boldsymbol{A}) \tag{1.62}$$

$$\frac{\mathrm{d}L}{\mathrm{d}\boldsymbol{\gamma}} = \boldsymbol{A}\boldsymbol{w} - \boldsymbol{b} = \boldsymbol{0}, \quad \boldsymbol{A}\boldsymbol{w} = \boldsymbol{b} \tag{1.63}$$

Inserting Equation 1.62 into Equation 1.63 and solving the resulting equation for the Lagrange multipliers, we arrive at

$$\lambda\boldsymbol{A}\boldsymbol{\Omega}^{-1}\boldsymbol{\mu} - \boldsymbol{b} = \lambda\boldsymbol{A}\boldsymbol{\Omega}^{-1}\boldsymbol{A}'\boldsymbol{\gamma}$$

$$\boldsymbol{\gamma} = \frac{\boldsymbol{A}\boldsymbol{\Omega}^{-1}\boldsymbol{\mu}}{\boldsymbol{A}\boldsymbol{\Omega}^{-1}\boldsymbol{A}'} - \frac{1}{\lambda}\frac{\boldsymbol{b}}{\boldsymbol{A}\boldsymbol{\Omega}^{-1}\boldsymbol{A}'} \tag{1.64}$$

Substituting Equation 1.64 into Equation 1.62, we finally get the optimal solution under linear equality constraints

$$\boldsymbol{w}^* = \boldsymbol{\Omega}^{-1}\boldsymbol{A}'(\boldsymbol{A}\boldsymbol{\Omega}^{-1}\boldsymbol{A}')\boldsymbol{b} + \lambda\boldsymbol{\Omega}^{-1}(\boldsymbol{\mu} - \boldsymbol{A}'(\boldsymbol{A}\boldsymbol{\Omega}^{-1}\boldsymbol{A}')^{-1}\boldsymbol{A}\boldsymbol{\Omega}^{-1}\boldsymbol{\mu}) \tag{1.65}$$

The optimal solution is split into a (constrained) minimum-variance portfolio and a speculative portfolio. This is known as "two-fund separation", and can be seen from Equation 1.65, where the first term depends neither on expected returns nor on risk tolerance – and is hence the minimum-risk solution – whereas the second term is sensitive to both inputs. Constrained optimisation reduces the efficiency of the solution as a constrained solution must be less optimal than an unconstrained solution. The loss in efficiency can be measured as the difference between a constrained and an unconstrained solution.[62]

However, it should be stressed that not every difference is for real, ie, statistically or economically significant.[63] We would, for example, expect optimising across sectors to yield better results than optimisation across countries simply because there are more sectors to choose from.

To test for significance, we use the Sharpe ratio, SR. Consider the simple case of running an unconstrained optimisation with k' assets and a constrained optimisation with only k assets $(k' > k)$. Whether the better performance of the unconstrained portfolio is statistically significant can be calculated using (Glen and Jorion 1993)

$$\frac{(T-k')(k'-k)(\mathrm{SR}'^2-\mathrm{SR}^2)}{1+\mathrm{SR}^2} \sim F_{k',T-(k'+k+1)} \tag{1.66}$$

where T is the number of observations and SR$'$ is the Sharpe ratio of the unconstrained strategy. The above statistic is F-distributed. However, as we will see in Chapter 3, this is not the only technique for testing whether the performance of two portfolios is significantly different.

APPENDIX C: FACTOR RISK CONTRIBUTIONS

So far we have assumed that we cannot decompose the uncertainty in asset returns into common factors. However, stocks are at least partly driven by characteristics (industry, country, size, etc) which they share with many other stocks. Suppose that we can write the risk premium of a given stock as a combination of these factor returns weighted by their respective factor exposures[64]

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

where $\boldsymbol{r}$ is a $k \times 1$ vector of risk premiums (asset return minus cash), $\boldsymbol{X}$ is a $k \times p$ matrix of factor exposures (sometimes called "loadings", particularly if they are derived from a principal component analysis), $\boldsymbol{f}$ is a $p \times 1$ matrix of factor returns and $\boldsymbol{u}$ is a $k \times 1$ vector of asset-specific returns that are both uncorrelated with factor returns and uncorrelated across assets as they are specific to a particular company.[65] The covariance matrix of excess returns can then be expressed as[66]

$$\begin{aligned}
E(\boldsymbol{r}\boldsymbol{r}') &= E(\boldsymbol{X}\boldsymbol{f}+\boldsymbol{u})(\boldsymbol{X}\boldsymbol{f}+\boldsymbol{u})' \\
&= E(\boldsymbol{X}\boldsymbol{f}\boldsymbol{u}') + E(\boldsymbol{X}\boldsymbol{f}\boldsymbol{f}'\boldsymbol{X}') + E(\boldsymbol{u}\boldsymbol{u}') + E(\boldsymbol{u}\boldsymbol{X}'\boldsymbol{f}) \\
\boldsymbol{\Omega} &= \boldsymbol{X}\boldsymbol{\Omega}_{ff}\boldsymbol{X}' + \boldsymbol{\Omega}_{uu}
\end{aligned}$$

where $\boldsymbol{\Omega}_{ff}$ denotes the $p \times p$ covariance matrix of factor returns and $\boldsymbol{\Omega}_{uu}$ is a $k \times k$ covariance matrix (diagonal matrix) of asset-specific returns. We can now decompose portfolio risk into a common and a specific part[67]

$$\sigma_{\mathrm{p}}^2 = \boldsymbol{w}'\boldsymbol{X}\boldsymbol{\Omega}_{ff}\boldsymbol{X}'\boldsymbol{w} + \boldsymbol{w}'\boldsymbol{X}\boldsymbol{\Omega}_{uu}\boldsymbol{w}$$

Using the same logic as before, we get for the marginal factor contribution to risk (MFCTR)

$$\mathbf{MFCTR} = \frac{\mathrm{d}\sigma_{\mathrm{p}}}{\mathrm{d}(\boldsymbol{X}'\boldsymbol{w})} = \frac{\boldsymbol{\Omega}_{ff}\boldsymbol{X}'\boldsymbol{w}}{\sigma_{\mathrm{p}}}$$

where **MFCTR** is an $f \times 1$ vector. The calculations introduced in this appendix will be applied in Chapter 8.

EXERCISES

1. Assume you are given an investment universe of k securities: all assets exhibit equal correlation, equal volatility and equal expected excess returns.

(a) What does the volatility of an equally weighted portfolio converge to as the number of assets goes to infinity?

(b) What does the portfolio Sharpe ratio converge to?

(c) What is the lower limit on correlation between assets as a function of k? Is there an upper limit? Can you explain the difference?

2. Use matrix algebra to find the maximum Sharpe ratio portfolio (ie, the portfolio that offers the highest return per unit of risk. Show that the Sharpe ratio is linear homogeneous. What does this mean economically?

3. Calculate expected exposure, variance and characteristic exposure per volatility to an arbitrary characteristic portfolio, $\boldsymbol{w}_{\zeta} = (\boldsymbol{\xi}'\boldsymbol{\Omega}^{-1}\boldsymbol{\xi})^{-1}\boldsymbol{\Omega}^{-1}\boldsymbol{\xi}$.

4. Assume we need to constrain the optimal portfolio to be neutral to any given characteristic $\boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is an $n \times 1$ vector of asset characteristics. For example, if we require beta neutrality, it simply becomes a vector of betas while for cash neutrality it becomes a vector of 1s. The portfolio optimisation problem now becomes

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu} - \frac{1}{2\lambda}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{w} - \upsilon(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\xi})$$

Solve for the optimal weight vector. What is the value of the Lagrange multiplier?

5. You hold three uncorrelated securities with volatilities of 20%, 10% and 5% in proportions of 50%, 20% and 30% in your portfolio. Your return expectations (excess returns over cash) for these assets have been 10%, 8% and 2%. Calculate the minimum variance portfolio. Compare the marginal contributions to risk for each asset. What do you find? Calculate the maximum Sharpe ratio portfolio. Calculate the percentage contributions to risk and the individual asset betas relative to the maximum Sharpe ratio portfolio. What do you find? Compare your portfolio with linear combinations of minimum variance and maximum Sharpe ratio portfolio. What do you find and what does this tell you about the efficiency of your portfolio?

6. Which of the following is true?

(a) Portfolio variance is linear homogeneous.

(b) The percentage contribution to risk is the same for all assets in the minimum variance portfolio.

(c) The percentage contribution to risk equals portfolio equals the portfolio weight in the minimum risk portfolio.

(d) Portfolio constraints will always improve *ex ante* efficiency.

(e) Portfolio constraints will always worsen *ex post* efficiency.

(f) Highly volatile assets will always increase portfolio risk.

(g) Excess returns (versus cash) are real returns.

(h) The Sharpe ratio of all assets will be the same in equilibrium.

[1] The reader is assumed to be familiar with the basics of mean–variance optimisation. A concise review can be found in Elton and Gruber (1995). This chapter will focus on selected issues not commonly found in textbooks.

[2] Discussions of the theory of utility can be found in Gollier (2001). Quadratic utility is of the form $u(w) = w - \frac{1}{2}bw^2$, where $u(w)$ expresses utility, a measure of happiness, as a function of uncertain wealth, w. In this example greater utility is not strictly an increase in wealth as, depending on the value of b, utility will decline for large enough w. Thus, investors will not always prefer more to less. Moreover, richer investors hold a smaller fraction in risky assets, which is not always plausible (increasing relative risk aversion).

[3] Returns are multivariate normal if they are jointly normal. Normality for a single return series is not enough for multivariate normality. Normality itself is a convenient feature as it allows returns to be easily aggregated across time and across holdings.

Conceptually, returns cannot be normally distributed as they are naturally truncated at −100% (maximum loss). We will, however, ignore this "technicality", keeping very much in line with most of the literature on portfolio selection. Focusing instead on log returns (which are normally distributed) would add the complication that log returns are not additive (the weighted sum of log returns is not the portfolio log return). More importantly, log returns offer very little additional insight into portfolio construction issues.

4 One common method is to use the average (local) risk premium as input. The average return characterises the expected return, while the median return does not. Working with the risk premium (risky security return minus the risk-free rate) has the advantage that it is more stable than pure returns as it takes out at least part of the interest rate component in returns. However, it is necessary to define what is meant by risk-free rate. If we focus on nominal capital growth, the risk-free rate coincides with the zero rate for the corresponding time horizon (it is risk-free because it is known at the beginning of the time horizon). If instead we focus on real capital growth, the risk-free security becomes inflation-linked bonds, which guarantee a real rate. A further advantage of the local risk premium approach is that it can easily be converted into different home currencies by adding on the domestic cash rate (domestic cash rate plus local risk premium equals foreign return hedged into home currency).

5 This book follows the convention that vectors and matrixes are represented by bold symbols. Estimation errors are dealt with in later chapters (Chapters 4 and 5 in particular).

6 Portfolio diversification lies at the heart of portfolio theory. Adding another asset to a portfolio (keeping total investments constant) will reduce portfolio risk. Suppose that an individual stock return can be written as market return plus stock-specific return ($R_i = R_\mathrm{m} + \alpha_i$) and that k equally weighted stocks are combined into one common portfolio. The risk (variance) of this portfolio can be calculated as

$$\operatorname{var}\left(\sum_{i=1}^{k} w_i R_i\right) = \left(\sum_{i=1}^{k} \frac{1}{k} R_i\right) = \left(\sum_{i=1}^{k} \frac{1}{k}(R_\mathrm{m} + \alpha_i)\right) = \operatorname{var}(R_\mathrm{m}) + \frac{\operatorname{var}(\alpha)}{k}$$

Adding more assets will result in a portfolio that carries only market risk, with stock-specific risk diversified away. This is also called "micro-diversification". However, groups of stocks are exposed to common macroeconomic factors, and little research has been done on *macro*-diversification, ie, understanding the returns of different asset classes in different economic environments. Questions such as whether hedge funds or commodities are the better hedge against down markets are rarely addressed but would give insight into macro-diversification.

7 The "≈" symbol has been used rather than "=" to emphasise that Equation 1.4 might hold approximately for a wide range of utility functions. Section 1.2 reviews the approximating properties of Equation 1.3.

8 This section will deal with linear equality constraints rather than inequality constraints as little is gained by extending the analysis to constraints that are mathematically more demanding but offer no additional investment insight as commercial solutions are now readily available.

9 Note that $\mathrm{d}(\boldsymbol{x}'\boldsymbol{\Omega}\boldsymbol{x})/\mathrm{d}\boldsymbol{x} = (\boldsymbol{\Omega} + \boldsymbol{\Omega}')\boldsymbol{x}$, which equals $2\boldsymbol{\Omega}\boldsymbol{x}$ if $\boldsymbol{\Omega} = \boldsymbol{\Omega}'$, ie, if $\boldsymbol{\Omega}$ is symmetric. See Green (2000, p. 52) for more on this.

10 Fama and MacBeth (1973).

11 For an excellent treatment of testing quantitative factors, see Sorensen *et al* (2007).

12 Note that $(\boldsymbol{\xi}'\boldsymbol{\Omega}^{-1}\boldsymbol{\xi})^{-1}$ is just a scaling factor aimed at achieving the right risk level, but it keeps relative allocations "untouched".

13 See Ziemba and Mulvey (1998) for a detailed presentation of many asset–liability management issues.

14 Private investors, for example, need to think about liabilities as a real consumption stream after retirement.

15 Writing surplus returns as change in surplus relative to assets rather than relative to surplus has the advantage that surplus returns are always defined (even for zero surplus) and do not become too large (otherwise a surplus rising from 1% to 2% would lead to a 100% surplus return).

16 However, we do not assume f to be fixed (known at start-of-period) as real asset–liability management would also change liabilities (sale, reinsurance, etc).

17 One of the difficulties in asset–liability management arises from the non-existence of a liability-mimicking asset that, if bought, would hedge out all the liability risks completely. This is often

caused by inflation-indexed liabilities or final wage-related schemes, which create unhedgable equity-linked liabilities. Pension funds will always have to accept some liability noise.

18 Each entry in $\boldsymbol{\Omega}_{\text{surplus}}$ can be expressed as $\operatorname{cov}(R_i - f R_l, R_j - f R_l)$.

19 Alternatively, we can define the vector of covariances between assets and our liability as $\boldsymbol{\Gamma} = [\sigma_{1l} \cdots \sigma_{kl}]'$. As the funding ratio, liability return and liability risk are fixed (known at the start of the investment period), we are left to maximise $\boldsymbol{w}'\boldsymbol{\mu} - \frac{1}{2}\lambda(\boldsymbol{w}'\boldsymbol{\Omega}\boldsymbol{w} - 2f\boldsymbol{w}'\boldsymbol{\Gamma})$ subject to constraints. Sharpe and Tint (1990) call the term $2f\boldsymbol{w}'\boldsymbol{\Gamma}$ "liability hedging credit" as assets that have positive correlation with liabilities add positively to an investor's utility. If the value of assets becomes very high relative to liabilities, the optimisation problem looks increasingly like an asset-only optimisation as f becomes small and there is little effect from the liability-hedging credit.

20 In a simple two-asset case (with the adding-up constraint that both assets have to sum to one), the solution with the smallest surplus volatility is

$$w_1 = \frac{\sigma_{22} - \sigma_{12} + f(\sigma_{1l} - \sigma_{2l})}{\sigma_{11} + \sigma_{22} - 2\sigma_{12}}$$

If $\sigma_{1l} = \sigma_{2l}$, we arrive at the same solution as without liabilities. A special case of this is $\sigma_{1l} = \sigma_{2l} = \mathbf{0}$, which applies to cash-like liabilities.

21 See Chapter 8 on benchmark-relative optimisation.

22 The entry of 0.04 indicates that asset one is characterised by 20% volatility (square root of 4% variance). This means that two-thirds of all return observations lie between the mean return plus or minus two standard deviations, ie, ±20%.

23 The constrained frontier touches the unconstrained frontier where all portfolio weights sum to one.

24 Strictly, investors with quadratic utility would not prefer more wealth to less wealth as utility falls after the satiation point. Quadratic utility also induces increasing absolute risk aversion (richer individuals hold smaller amounts in risky assets than less rich individuals).

25 Huang and Litzenberger (1988, p. 60).

26 Levy and Markowitz (1979).

27 Favre and Galeano (2001) and Fung and Hsieh (1997).

28 This is only true if returns are normally distributed. In the case of non-normal returns, the maximum-growth portfolio might not lie on the efficient frontier as positive skewness adds to long-term growth whereas kurtosis detracts from it.

29 This is also called "Kelly betting". See Kelly (1956).

30 See Mossin (1968). Kritzman (2000) also provides an excellent review.

31 Time-varying investment opportunities and their impact on the riskiness of strategic asset allocation decisions are discussed in Campbell and Viceira (2002).

32 Barberis (2000) and Xia (1999).

33 For an accessible introduction readers are referred to Brandimarte (2001).

34 We use the term institutional investors as synonymous to corporate plan sponsors or insurance companies.

35 Litterman (1996) coined the terms "hot spots" (areas where most of the portfolio risks are concentrated) and "best hedges" (risk-producing positions).

36 Grinold and Kahn (2000) is a complete and highly recommended source on the decomposition of portfolio risk.

37 Maximum diversification is sometimes thought to be achieved with the "equal risk portfolio".

38 Readers are encouraged to repeat all calculations for themselves. The book contains the information required for almost all examples in this and the following chapters.

[39] Bayer (1998).

[40] Jorion (2001).

[41] McCarthy (2000, p. 103).

[42] Chow and Kritzman (2001, p. 58).

[43] De Bever *et al* (2000, p. 283).

[44] Jorion (2001).

[45] VaR has come under attack from Artzner *et al* (1999), who found that as one deviates from the normality assumption on returns, VaR is not necessarily sub-additive, ie, the VaR of a portfolio might be higher than the combined VaRs of the single positions.

[46] The covariance matrix is, in most applications, estimated either as a multifactor risk model or as a sample covariance matrix (which is ill-conditioned when the number of assets becomes large relative to the number of observations and of limited use in optimisation as it contains random noise as well as significant relationships, but it nevertheless provides a maximum likelihood forecast of risk). The greatest advantages of using the covariance matrix are the reduction of dimensions (low number of factors relative to number of assets), the stability of co-movements (as noise is separated from systematic relations) and the intuitive understanding it affords of the key drivers of portfolio risk (at least for economic multifactor models). Factors are derived either with the use of economic theory (macroeconomic or fundamental factors) or purely on statistical grounds. While statistical models (principal components, etc) are best for portfolio optimisation with regard to passive management, economically based factor models can also serve to identify (and neutralise or sharpen) groups of bets that are linked by common economic factors. The way risk is measured and decomposed has to coincide with the way a risk allocation is decided on; hence, risk management and return forecasting models should not look too different.

[47] Johnston (1984, p. 150).

[48] The eigenvalues can also be used to calculate the common variance that is explained by so-called "principal components" (uncorrelated artificial variables that are constructed to maximise explanatory power). The ith component explains $e_i / \sum e_i$ of the variance in the three variables (note: eigenvalues are ordered starting with the biggest). In this case the first component explains about 87% of the variance, which directs us towards a strong common trend. (All matrix-orientated programming languages provide routines for calculating eigenvalues. Readers can also use PopTools by Greg Hood.)

[49] For a symmetric 3 × 3 matrix all correlation coefficients must satisfy

$$\rho_{12}^2 + \rho_{13}^2 + \rho_{23}^2 - 2\rho_{12}\rho_{13}\rho_{23} \leqslant 1$$

to describe a valid relationship.

[50] Note that this is a rather technical correction with no relation to the data. In general we would like to use the information available in past returns. See Ledoit (1997) for a Bayesian approach.

[51] See Sharifi *et al* (2004) for a review in the context of portfolio optimisation.

[52] Edelman (1988).

[53] Alternatively, we can bootstrap the small sample distribution of $e_{\max}$ rather than relying on the large sample equivalent.

[54] Malevergne and Sornette (2004) and Kwapien *et al* (2006).

[55] Suleman *et al* (2006).

[56] Stevens (1998).

[57] Alternatively, we could incorporate all other items (tax revenues, government spending, pension liabilities, etc) into the government budget surplus/deficit calculation. However, given the relatively low volatility of these positions, we will ignore this problem. It will not change the nature of our findings.

[58] Chapters 4 and 5 deal with estimation error in greater detail.

59 Estimation error is given by $\sigma/\sqrt{T} = 0.2/\sqrt{T}$, and the confidence interval is calculated as

$$\left[+ \frac{\sigma}{\sqrt{T}} z_\alpha - z_\alpha \frac{\sigma}{\sqrt{T}} \right]$$

where z_α denotes the critical value of a standard normal distribution. In this case $z_\alpha = 1.96$. The sampling error on the annual mean estimate depends only on the length of the data period (in years) and not on the data frequency itself.

60 Campbell *et al* (1997, p. 364) show that

$$\operatorname{var}(\hat{\sigma}^2) = \left(\frac{T}{\Delta t} - 1 \right)^{-1} 2\sigma^2$$

61 An adding-up constraint would be written as $[1 \cdots 1]\boldsymbol{w} = \sum w_i = 1$.

62 However, constraints are sometimes justified as safeguards against bad inputs. We will review this issue in Chapter 4.

63 The reader is referred to the literature on mean–variance spanning tests, which considers whether a new asset added to the investment universe provides an investment opportunity or whether it is spanned (priced) by already existing assets. See Cochrane (2001).

64 Factor models work well on individual stocks, but they hardly add anything to indexes as indexes are already very diversified portfolios.

65 Essentially this means that $E(\boldsymbol{X f u'}) = \mathbf{0}$.

66 Note that $(\boldsymbol{A} + \boldsymbol{B})' = (\boldsymbol{B}' + \boldsymbol{A}')$ and $(\boldsymbol{AB})' = (\boldsymbol{B}'\boldsymbol{A}')$.

67 The same can be done for individual betas

$$\boldsymbol{\beta} = \frac{\boldsymbol{\Omega w}}{\boldsymbol{w}'\boldsymbol{\Omega w}} = \frac{\boldsymbol{X \Omega}_{ff} \boldsymbol{X}' \boldsymbol{w}}{\boldsymbol{w}'\boldsymbol{\Omega w}} + \frac{\boldsymbol{w}'\boldsymbol{\Omega}_{uu}\boldsymbol{w}}{\boldsymbol{w}'\boldsymbol{\Omega w}}$$

REFERENCES

Almgren, R., and N. Chriss, 2006, "Optimal Portfolios from Ordering Information", *Journal of Risk* 9(1), pp. 1–22.

Artzner, P., F. Delbaen, J. Eber and D. Heath, 1999, "Coherent Measures of Risk", *Mathematical Finance* 9, pp. 203–28.

Barberis, N., 2000, "Investing for the Long Run when Returns Are Predictable", *Journal of Finance* 55, pp. 225–64.

Bayer, K., 1998, "Vom traditionellen Management zur Portfolio Factory: Anforderungen und Ziele", in C. Kutscher and G. Schwarz (eds), *Aktives Portfoliomanagement* (Zürich: Neue Züricher Zeitung).

Brandimarte, P., 2001, *Numerical Methods in Finance* (New York: John Wiley & Sons).

Campbell, J., and L. Viceira, 2002, *Strategic Asset Allocation: Portfolio Choice for Long-Term Investors* (Oxford University Press).

Campbell, J., A. Lo and C. MacKinlay, 1997, *The Econometrics of Financial Markets* (Princeton, NJ: Princeton University Press).

Chow, G., and M. Kritzman, 2001, "Risk Budgets", *Journal of Portfolio Management* 27, Winter, pp. 56–60.

Cochrane, J., 2001, *Asset Pricing* (Princeton, NJ: Princeton University Press).

De Bever, L., W. Kozun and B. Zwan, 2000, "Risk Budgeting in a Pension Fund", in L. Rahl (ed), *Risk Budgeting* (London: Risk Books).

Edelman, A., 1988, "Eigenvalues and Condition Numbers of Random Matrices", *SIAM Journal* 9, pp. 543–60.

Elton, E., and M. Gruber, 1995, *Modern Portfolio Theory and Investment Analysis*, Fifth Edition (New York: John Wiley & Sons).

Fama, E. F., and J. MacBeth, 1973, "Risk Return and Equilibrium: Empirical Tests", *Journal of Political Economy* 81, pp. 607–36.

Favre, L., and J. Galeano, 2001, "Portfolio Allocation with Hedge Funds, Case Study of a Swiss Institutional Investor", Working Paper, UBS Warburg.

Fung, W., and D. Hsieh, 1997, "Is Mean Variance Analysis Applicable to Hedge Funds?", Working Paper, Duke University, North Carolina.

Glen, J., and P. Jorion, 1993, "Currency Hedging for International Portfolios", *Journal of Finance* 48, pp. 1865–86.

Gollier, C., 2001, *The Economics of Time and Uncertainty* (Cambridge, MA: MIT Press).

Green, W., 2000, *Econometric Analysis*, Fourth Edition (New York: Prentice-Hall).

Grinold, R., and R. Kahn, 2000, *Active Portfolio Management*, Second Edition (New York: McGraw-Hill).

Huang, C.-F., and R. H. Litzenberger, 1988, *Foundations for Financial Economics* (Englewood Cliffs, NJ: Prentice Hall).

Johnston, J., 1984, *Econometric Methods*, Third Edition (New York: McGraw-Hill).

Jorion, P., 2001, *Value at Risk*, Second Edition (New York: McGraw-Hill).

Kelly, J., 1956, "A New Interpretation of Information Rate", *Bell System Technical Journal* 35, pp. 917–26.

Kritzman, M., 2000, *Puzzles of Finance* (New York: John Wiley & Sons).

Kwapien, J., S. Drozdz and P. Oswiecimka, 2006, "The Bulk of the Stock Market Correlation Matrix Is Not Pure Noise", *Physica A* 359, pp. 589–606.

Ledoit, O., 1997, "Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection", Working Paper, University of California, Los Angeles.

Levy, H., and H. Markowitz, 1979, "Approximating Expected Utility by a Function of the Mean and Variance", *American Economic Review* 69, pp. 308–17.

Litterman, B., 1996, "Hot Spots and Hedges", *Journal of Portfolio Management*, Special Issue, pp. 52–75.

Malevergne, Y., and D. Sornette, 2004, "Collective Origin of Co-existence of Apparent Random Matrix Theory Noise and of Factors in Large Sample Correlation Matrices", *Physica A* 331, pp. 660–8.

McCarthy, M., 2000, "Risk Budgeting for Pension Funds and Investment Managers Using VaR", in L. Rahl (ed), *Risk Budgeting* (London: Risk Books).

Mossin, J., 1968, "Optimal Multiperiod Portfolio Policies", *Journal of Business* 41, pp. 215–29.

Sharifi, S., M. Crane, A. Shamaie and H. Ruskin, 2004, "Random Matrix Theory for Portfolio Optimization: A Stability Approach", *Physica A* 335, pp. 629–43.

Sharpe, W., and L. Tint, 1990, "Liabilities: A New Approach", *Journal of Portfolio Management* 16, pp. 5–11.

Sorensen, E., E. Qian and R. Hua, 2007, *Quantitative Equity Portfolio Management: Modern Techniques and Applications* (Chapman & Hall).

Stevens, G., 1998, "On the Inverse of the Covariance Matrix in Portfolio Analysis", *Journal of Finance* 53, pp. 1821–7.

Suleman, O., M. McDonald, S. Williams, S. Howison and N. F. Johnson, 2006, "Implications of Correlation Cleaning for Risk Management", URL: http://www.maths.ox.ac.uk/~mcdonal4.

Xia, Y., 1999, "Learning about Predictability: The Effect of Parameter Uncertainty on Dynamic Optimal Consumption and Asset Allocation", Working Paper, University of California, Los Angeles.

Ziemba, W., and J. Mulvey, 1998, *Worldwide Asset and Liability Modelling* (Cambridge University Press).

2

Application in Mean–Variance Investing

2.1 CLUSTERING TECHNIQUES AND THE INVESTMENT UNIVERSE

2.1.1 The correlation problem

Few investors are aware that the definition of the investment universe itself has a considerable impact on the outcome of portfolio construction. If, for example, the investment universe is constrained to government bonds and emerging market bonds, it is obvious that all combinations are efficient and investors are likely to place a considerable allocation in emerging market bonds. However, as soon as high-yield bonds and emerging market equities are introduced, this situation might change due to the correlation between these assets.

To increase the transparency of the asset allocation process as well as to avoid the accumulation of estimation errors, portfolio optimisation should be limited to groups of assets that have high intragroup and low intergroup correlations.[1] The benefit of this correlation-guided asset class definition is a weakened sensitivity of the optimal portfolio solution with respect to mean return estimates (which carry the highest estimation error, as can be seen from Appendix A of Chapter 1 (on page 39)). Again, we start from the definition of optimal portfolio weights, $\boldsymbol{w}^*$, applied to the two-asset case

$$\begin{aligned} \boldsymbol{w}^* &= \begin{bmatrix} w_1^* \\ w_2^* \end{bmatrix} \\ &= \lambda \begin{bmatrix} \boldsymbol{\Omega}_{11}^{-1} & \boldsymbol{\Omega}_{12}^{-1} \\ \boldsymbol{\Omega}_{21}^{-1} & \boldsymbol{\Omega}_{22}^{-1} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \end{aligned} \tag{2.1}$$

where $\boldsymbol{\Omega}_{11}^{-1}$ denotes element $(1, 1)$ of $\boldsymbol{\Omega}^{-1}$. Taking the first derivative with respect to the expected mean of the first asset, μ_1, and applying the rules for the inverse of a partitioned matrix, we get[2]

$$\begin{aligned}\frac{\mathrm{d}w_1^*}{\mathrm{d}\mu_1} &= \lambda\boldsymbol{\Omega}_{11}^{-1} = \lambda\frac{\sigma_{22}}{\sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21}} \\ &= \lambda\frac{1}{\sigma_{11} - \rho^2\sigma_{11}} \qquad (2.2)\end{aligned}$$

where ρ is the correlation between the two assets. As ρ approaches 1, portfolio weights will react very sensitively to changes in means; as assets become more similar, any extra return becomes increasingly important for the allocation decision. Portfolio optimisation with highly correlated assets will almost certainly lead to extreme and undiversified results.

2.1.2 Cluster analysis

To avoid the accumulation of estimation errors mentioned above, we need to identify groups, or "clusters", of assets on the basis of the correlation between individual assets. The technique for doing this is cluster analysis.[3] Clustering methods can be divided into partitioning methods, which optimally separate the k assets into K ($K \leqslant k$) distinct clusters, where K is given exogenously rather than derived within the analysis (producing a single partition), and hierarchical methods, which establish k partitions within one run. Hierarchical methods might appear to make partitioning methods obsolete as they determine all possible clusters in one run. However, this is achieved at the cost of less optimal clusters as hierarchical methods always look for the best add-on to an existing cluster and past add-ons are treated as fixed. We will focus here on hierarchical methods as they are not particularly computer-intensive and require less explanation. Readers wishing to arrive at a better grouping of the investment universe should experiment with partitioning methods (see note 3).

We start the grouping process with all assets still separate, each asset being viewed as a cluster of its own.[4] The first step is to merge the two "closest" assets, ie, those with the highest correlation. In clustering methodology this is also called the "minimum distance" and is measured as 1 − Correlation.[5] We will explain the algorithm using the data for seven hedge fund sub-indexes (monthly return data published by Hedge Fund Research, Inc, for the period January 1990 to March 2001).[6] The distance matrix is given in Table 2.1 and will allow the reader to follow through our example.

Table 2.1 Distance (1 − Correlation) matrix for hedge fund data

	CA	DD	LSI	EDS	MI	MT	SS
CA	0.00	0.33	0.49	0.35	0.60	0.63	1.38
DD	0.33	0.00	0.36	0.18	0.49	0.59	1.54
LSI	0.49	0.36	0.00	0.22	0.37	0.29	1.85
EDS	0.35	0.18	0.22	0.00	0.37	0.47	1.66
MI	0.60	0.49	0.37	0.37	0.00	0.45	1.46
MT	0.63	0.59	0.29	0.47	0.45	0.00	1.70
SS	1.38	1.54	1.85	1.66	1.46	1.70	0.00

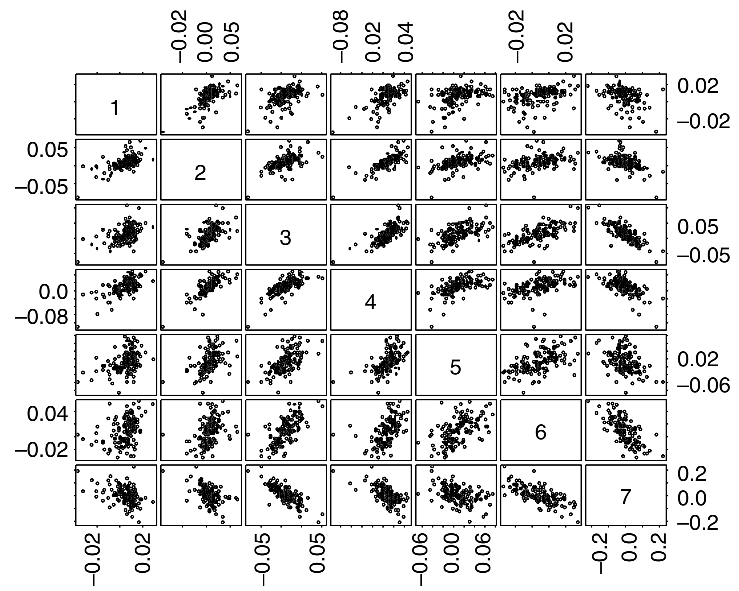
CA, convertible arbitrage; DD, distressed debt; LSI, long–short investing; EDS, event-driven strategies; MI, macro-investing; MT, market timing; SS, short sellers.

The greatest distance is that between short sellers and all other hedge funds. The reason for this is the substantial negative correlation of this hedge fund style with other styles, as can be seen from the scatter plots in Figure 2.1. The scatter clouds of short sellers are negatively sloped. The smallest distance, though, is that between "event-driven strategies" and "distressed debt" (which, arguably, is also event-driven). How do we proceed from here? We compare the distances between all remaining clusters, including the cluster of event-driven strategies and distressed debt, and calculate, pairwise, the distances between this newly formed cluster, C_1, and the other assets (for notational convenience also called clusters) according to

$$\text{Distance}(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{\substack{i \in C_1 \\ j \in C_2}} \text{Distance}(i, j) \qquad (2.3)$$

where $|C_n|$ is the number of objects (assets) in cluster n, the distance between two clusters is the average distance between their individual elements, and distance is measured as 1 − Correlation. Casual inspection of Table 2.1 would tempt us to conclude that, after distressed debt (0.18), long–short investing is the closest (0.22) to event-driven strategies and, hence, that we should add it to cluster C_1. However, as we have to calculate the average distance between two clusters and the distance between long–short investing and distressed debt is relatively high (0.36), we find that the next closest assets are long–short investing and market timing (0.29), indicating that long–short investors still have considerable directional market exposure.

Figure 2.1 Bivariate scatter plots



1, Convertible arbitrage; 2, distressed debt; 3, long–short; 4, event-driven strategies; 5, macro-investing; 6, market timing; 7, short seller.

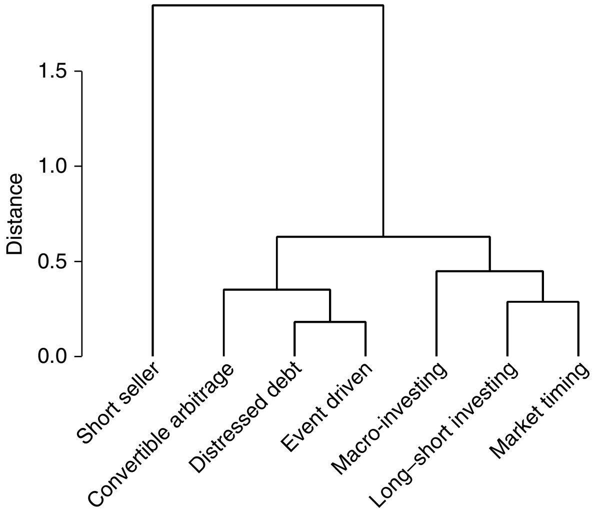
The complete picture is summarised in Figure 2.2, where the axis represents the distance between two clusters. With a bit of economic judgment we conclude that there are three clusters: strategies with directional market exposure (market timing, long–short investing and macro-investing); strategies with reverse market exposure (short sellers); and strategies that focus on company-specific events (convertible arbitrage, distressed debt and event-driven strategies). However, we should keep in mind that this partitioning is based on some marginal differences in the distance matrix, and in such cases an alternative algorithm may be desirable.[7]

2.2 ILLIQUID ASSETS: CORRECTING FOR AUTOCORRELATION

2.2.1 The problem

Some asset classes appear to show much less risk than practitioners commonly believe to be reasonable; for example, corporate high-yield indexes exhibit much less volatility than their hybrid structure between equity and fixed income might otherwise suggest.

Figure 2.2 Dendrogram for hedge fund data



For the period September 1986 to December 2000 the annualised volatility of the Merrill Lynch High Yield Master II Index was 5.6%.[8] The same problem arises with real estate, which shares the characteristics of fixed income (a steady stream of rental income) and equity (a claim on real assets) but, measured empirically, shows less than fixed-income volatility. Historical annualised volatility for UK real estate over the period January 1988 to October 2000 was 3.0% – hardly imaginable for an asset class that returned 10.3% per annum during the same period. Most prominently, hedge funds also appear to show little risk relative to traditional investments, which significantly boosts their Sharpe ratios.

If the risk of an asset class is underestimated, too much capital will be allocated to it and there will be a serious loss of portfolio efficiency. To avoid such problems we have to look at the causes behind what appear to be biases in risk. The volatility bias mentioned above – the very low volatility of UK real estate returns relative to the size of the returns – is due to the high positive autocorrelation of the period-by-period returns, ie, a positive return one month is likely to be followed by a positive return the next month. A series of such monthly returns shows less historical volatility than an uncorrelated series, where a positive return is more likely to be followed by a negative return. Equally, positively autocorrelated (or "trending") returns are inevitably accompanied by a greater level of risk as there is little diversification between period-by-period returns – as would be the case with uncorrelated returns – and end-of-period wealth becomes more dispersed and, hence, more risky.

But where does the autocorrelation in returns come from? Common to all illiquid asset classes is that most observed prices are not market prices due to infrequent trading in illiquid securities. This applies equally to real estate, small companies, and high-yield and hedge funds. Suppose that at time *t* only half of the securities in an index react to some bad news. This does not mean that the other assets are unaffected by the news; it may only mean that it is not reflected in their prices at time *t* as no trade in them has taken place. If these untraded assets are then traded at time $t + 1$, their prices will reflect the time-*t* information with a lag of one period and will therefore generate positive autocorrelation in the return series for the index as a whole. Apart from the resulting downward-biased volatility, we might also see too low a correlation with other asset classes. Hedge funds are an excellent example of an illiquid asset class where prices often reflect the valuation of traders rather than market prices.

2.2.2 A simple method

Most problems have many cures, so this book will focus on a method of checking and correcting for autocorrelation in return series that is easy to implement. Using a simple filter of the form[9]

$$r_t^* = \frac{1}{1 - a_1} r_t - \frac{a_1}{1 - a_1} r_{t-1} \tag{2.4}$$

we can create a new transformed return series, r_t^*, using the returns, r, at times t and $t - 1$. The coefficient a_1 is estimated from an autoregressive first-order (AR(1)) model

$$r_t = a_0 + a_1 r_{t-1} + \varepsilon_t \tag{2.5}$$

where a_0 and a_1 are the regression coefficients and ε_t is the error term. The procedure is as follows.

1. Estimate the autoregressive model in Equation 2.5.
2. Use the returns at times t and $t - 1$ to calculate a new, filtered return for time t.
3. Repeat step 2 for all observations.[10]

Table 2.2 Results for autocorrelation example: betas for selected hedge funds

Type of hedge fund	a_1	β_0	β_0^*	$\beta_0 + \beta_1 + \beta_2 + \beta_3$
Convertible arbitrage	0.55 (7.66)	0.09	0.22	0.25
Distressed debt	0.52 (6.86)	0.18	0.44	0.49
Event-driven strategies	0.28 (3.56)	0.29	0.38	0.38
Macro-trading	0.18 (2.10)	0.29	0.37	0.52

Positive autocorrelation is indicated by the Student's *t* values in parentheses; see text for explanation of other symbols.

2.2.3 Example

Returning to our hedge fund example, we will apply this methodology to return series for convertible arbitrage, distressed debt, event-driven strategies and macro-investing. The series we will use are monthly return data for the period January 1990 to December 2000, available, as before, from Hedge Fund Research.[11] The aim is to show how to arrive at more realistic beta estimates with respect to a core holding in US equities (here represented by a Morgan Stanley Capital International fund, MSCI USA). When data is non-synchronous (due to infrequent trading resulting from illiquidity), we are more likely to capture the real dynamics by running a regression of returns against both lagged and contemporaneous market returns[12]

$$r_{it} = \alpha + \beta_0 r_{\mathrm{m},t} + \beta_1 r_{\mathrm{m},t-1} + \cdots + \beta_l r_{\mathrm{m},t-l} + \varepsilon_t \tag{2.6}$$

where r denotes risk premiums, β_0 is the contemporaneous beta and β_l is the regression coefficient on the lth lag of market returns, $r_{\mathrm{m},t-l}$. An improved measure of beta can be found by adding the significant lags to the contemporaneous beta $(\beta_0 + \beta_1 + \cdots + \beta_l)$. Alternatively, we could use the filter technique seen in Equation 2.4 to unsmooth the return series and calculate a standard regression beta from

$$r_{it}^* = \alpha + \beta_0^* r_{\mathrm{m},t} + \varepsilon_t \tag{2.7}$$

The results of this exercise are given in Table 2.2. All four hedge fund series show significant positive autocorrelation (as indicated by the t values in brackets), while contemporaneous betas, β_0, are considerably smaller than those which include lagged betas (fifth column of table). As expected, we find that contemporaneous betas on filtered series (fourth column) are significantly higher than those from ordinary regressions and are, in most cases, close to betas derived from regressions using lagged market returns. The betas from ordinary return regressions appear to underestimate the true market exposure. This feature is particularly troublesome as it overstates the diversifying properties of hedge funds. It also casts doubt on the standard practice of "mapping" implicit hedge fund market exposure through a regression of hedge fund returns on underlying asset class returns. Hence, it would be prudent to check assets that are suspected to be illiquid for autocorrelation and use the filtered estimates from Equation 2.7 for further analysis.

2.3 COVARIANCE IN GOOD AND BAD TIMES

Unfortunately, in times of market meltdown, just when portfolio managers need them most, correlations within an asset class increase.[13] This section will not attempt to forecast the change in input parameters, but instead will discuss tools for evaluating the diversifying properties of different assets in unusual times.

Is the low correlation in a full-sample covariance matrix (one that uses all available data) just an artefact of reasonably positive correlation in normal times, ie, most of the time, but of highly negative correlation in unusual times? Or is it evidence of a truly diversifying asset? As regulators (and supervisory boards) become increasingly concerned about short-term performance, investors often do not have the luxury of betting on average correlation. This is certainly true for most pension funds. The average correlation between bonds and equities is positive (equity markets rise when yields fall and bond returns rise too), but in times of crisis it becomes negative – government bonds outperform as investors move into the safe haven of high-grade fixed income. This will raise the liabilities of pension funds while at the same time massively reducing the assets of equity-orientated schemes. Surplus risk-based measures (see Section 2.1) calculated from average covariances will fail to spot these risks.

2.3.1 A statistical definition of unusual times

To come up with correlation and volatility estimates for normal and unusual times, we must first define "unusual times". We will define them according to their statistical distance from the mean vector, as follows[14]

$$(\boldsymbol{r}_t - \hat{\boldsymbol{\mu}}_0)'\hat{\boldsymbol{\Omega}}_0^{-1}(\boldsymbol{r}_t - \hat{\boldsymbol{\mu}}_0) = \boldsymbol{d}_t'\hat{\boldsymbol{\Omega}}_0^{-1}\boldsymbol{d}_t \tag{2.8}$$

where the distance vector $\boldsymbol{d}_t$ is a $k \times 1$ vector at time t, $\boldsymbol{r}_t$ is a vector of return observations for k assets at time t, $\hat{\boldsymbol{\mu}}_0$ is a $k \times 1$ vector of average returns and $\hat{\boldsymbol{\Omega}}_0$ is the unconditional covariance matrix (over all $t = 1, \dots, T$ observations). We calculate Equation 2.8 for each cross-section of stock returns and compare the distance vector with the critical value of $\chi^2(k)$. Note that this assume we exactly know what regime a return has been drawn from. This conflicts with more advanced Bayesian models where we assume data is drawn from a mixture of normal distributions, but update the probability of a given regime using the Bayes law. If we define an unusual observation as the outer 10% of a distribution (it can also be called the "outlier") and we look at five return series, our cut-off distance is 9.23. In Equation 2.8 the return distance is weighted by the inverse of the covariance matrix. This means that we take into account asset volatilities (the same deviation from the mean might be significant for low-volatility series but not necessarily for high-volatility series) as well as correlations (return difference of the opposite sign for two highly correlated series might be more unusual than for a series with negative correlation). Hence, in theory, outliers are not necessarily associated with down markets (although they often are in practice).[15]

2.3.2 An application to real data

As before, an example should help us to understand the mechanics. We use Morgan Stanley Capital International (MSCI) data for US and Japanese equities, Salomon Brothers data for medium-term US bonds (World Government Bond Index (WGBI)), and Hedge Fund Research, Inc. (HFR), hedge fund series for market-neutral and market-timing hedge funds. All data is monthly returns in US dollars (September 1990–March 2001). Using the methodology explained above, we can split the data into two regimes (normal times and unusual times, where unusual times are defined as market movements that only happen in 90% of all cases) and calculate correlations as well as volatilities for both regimes.

The correlations are given in Table 2.3, which shows that market-neutral hedge funds have little correlation with equity markets in normal times (although the correlation indicates that some of the funds in the index universe use some market leverage to create their

Table 2.3 Correlation in normal and unusual times

	MSCI US	**MSCI Japanese**	**SB WGBI**	**HFR neutral**	**HFR timing**
Normal times					
MSCI US	1.00	0.35	0.34	0.25	0.69
MSCI Japanese	0.35	1.00	0.08	0.05	0.48
SB WGBI	0.34	0.08	1.00	0.23	0.22
HFR neutral	0.25	0.05	0.23	1.00	0.23
HFR timing	0.69	0.48	0.22	0.23	1.00
Unusual times					
MSCI US	1.00	0.41	−0.18	0.39	0.44
MSCI Japanese	0.41	1.00	−0.09	0.39	0.04
SB WGBI	−0.18	−0.09	1.00	−0.25	−0.10
HFR neutral	0.39	0.39	−0.25	1.00	0.46
HFR timing	0.44	0.04	−0.10	0.46	1.00

HFR, Hedge Fund Research, Inc (market-neutral investing and market timing); MSCI, Morgan Stanley Capital International (US and Japanese equities); SB, Salomon Brothers (World Government Bond Index).

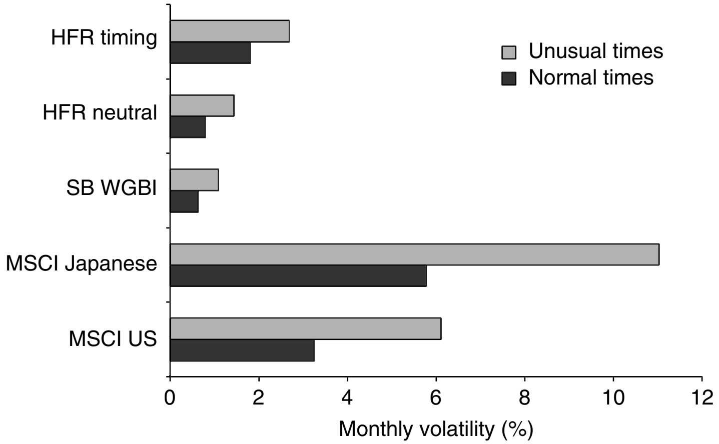
returns), but that they show increased dependency in unusual times. Conversely, market timers show a remarkable ability to decouple in unusual times. The table also confirms the crisis character of government bonds (negative correlation in times of crisis) as well as the stable–diversifying properties of Japanese equities.

For those investors concerned about unusual times, we could build a new covariance matrix as a mixture of both covariance matrixes with increasing weight put on the covariances in unusual times. To do so we need not only the objective probabilities from the correlation matrix for normal times, ie, 75% if we define outliers as those values that only occur in 25% of all cases) but also the relative risk aversion to both regimes. It can be shown that the new combined covariance matrix is a mixture of covariances in good and bad times

$$\boldsymbol{\Omega}_{\text{new}} = p\lambda_{\text{normal}}\,\boldsymbol{\Omega}_{\text{normal}} + \lambda_{\text{unusual}}(1-p)\boldsymbol{\Omega}_{\text{unusual}} \tag{2.9}$$

where p reflects the objective probabilities and λ_{normal} and λ_{unusual} are rescaled risk aversions such that their sum equals the actual risk aversion of an investor.

Figure 2.3 Volatility in normal and unusual times



Data sources as in Table 2.3.

Figure 2.3, which plots published volatility data for the investments in Table 2.3, confirms that the volatility for unusual observations is higher than for normal observations. Although the above analysis is useful as an exploratory tool, it should be noted that the results are sensitive to the inclusion of new assets, which might considerably change the properties of the already existing assets by changing previously normal periods to unusual periods and vice versa. This is why some researchers choose to define unusual periods in the narrow sense of down markets with particular reference to a prespecified core asset.[16]

2.4 THE CONDITIONAL COVARIANCE MATRIX

The goal of a (variance-based) risk model is to forecast the covariance matrix as efficiently and accurately as possible. So far we have been focusing on one of the simplest models of risk – the unconditional covariance matrix. Simplicity, however, comes at a price. We need $k(k-1)/2$ parameters to estimate (k variances on the main diagonal and the remaining upper or lower triangular half). Imagine an investment universe with $k = 2{,}000$ assets. We would need at least $2{,}000 \times 1{,}999/2 \approx 2$ million data points for our matrix to be positive semi-definite (ie, to be of rank 2,000). The larger the number of assets (relative to the number of observations), the more severe estimation error (low condition number) and portfolio instability become. That is where we need to impose a (factor) model on the data. The idea (hope) of imposing a parametric model is to reduce estimation error. However, we will only benefit from its structure as long as it is "halfway" correct. This section will lay out some intuition about factor models, but readers interested in this particular subject would be well advised to read the excellent text by Connor *et al* (2010).

In general, a factor model consists of factor loadings (betas, exposures) and factor realisations (factor returns), that allow us to link factor volatility to individual stocks and aggregate factor risks within portfolios. Different factor models vary on how factor loadings and factor returns are estimated and/or prespecified. Note that managing factor risk is extremely important as factor risk tends to be both leptokurtic and trending while residual risk tends to diversify away quickly (for a "good" factor model). We use factor models as they try to reduce estimation error by imposing structure and focus on the meaningful correlations rather than on noise. A generic model could look like

$$r_{it} = \alpha_i + \beta_{1i} \cdot f_{1t} + \beta_{2i} \cdot f_{2t} + \cdots + \beta_{Ki} \cdot f_{Kt} + \varepsilon_{it} \tag{2.10}$$

where r_{it} denotes the (total) return of stock i at date t, β_{1i} stands for the (time-invariant) factor exposure of stock i to factor 1, f_{1t} is the corresponding factor return and ε_{it} captures the risk unexplained in this regression, ie, how much of the stock's variation cannot be explained by factors. Probably the oldest and most intuitive factor models are macroeconomic models. These are time series models with prespecified factor returns, ie, the time series of factor returns in Equation 2.10 are given. They could be equity market or industry sector returns as well as innovations in oil price interest and inflation changes. When we run the above regression we only need to estimate factor exposures; factor returns are given. While clearly beta exposures are measured (estimated) with error, this measurement error does little harm at the portfolio level as it simply diversifies away when we add up individual beta exposures. The disadvantage with these models is that they leave too much open for explanation, or, to put it more bluntly, stocks are more dynamic than inflation rates. The most popular risk models are fundamental models as popularised by BARRA. They are inherently cross-sectional models and start with prespecified factor exposures to then estimate factor returns instead. Here lies a critical difference. If exposures are measured with error, the estimation error "creeps" into estimates of factor returns. However, once a factor return is contaminated by estimation error, there is no mechanism to diversify this error away. If a factor return is too high or too low, it will affect total portfolio risk. Note that factor returns can be driven considerably by outliers, which is why some software (Axioma) providers use robust regressions for Equation 2.10. Cross-sectional models seem better suited if your objective is to explain differences in return rather than risk. Finally, some investors use statistical models. These are time series models that simultaneously estimate factor returns and exposures. While they have a good in sample fit, they are more challenged with bad out-of-sample performance. Their biggest advantage as well as disadvantage is that they use few economic priors to identify factor risks. Hence, these statistical models might as well come from Mars. While this might look like a good idea, if we model an unknown emerging market with little fundamental data, it seems like a waste of knowledge where priors are widely available. After all, we know that company A is a biotech company, while company B is a pharmaceutical company. Statistical factors (also called blind factors for that reason) might need many data points to establish that they belong to different industries.

2.5 A SIMPLE MODEL FOR LIFE-CYCLE INVESTING

So far we have assumed that investors possess only financial wealth, an assumption that was necessary to guarantee that the optimal equity allocation was independent of our time horizon. This assumption will now be relaxed.

Life-cycle investing refers to the evolution of asset classes in an investor's portfolio as they approach retirement, and life-cycle concepts are becoming increasingly important as the asset management industry experiences a shift to defined-contribution plans. We know from Section 1.2 that if investors show constant relative risk aversion, the proportion of their wealth invested in risky assets will stay the same independent of total wealth, and if investment opportunities are constant (equities are not a hedge against bad times), the optimal equity allocation will be independent of the investment horizon.

In that case the optimal equity weight at time $t, w^*_{\text{equity},t}$, becomes constant: $\bar{w}_{\text{equity}}$.[17]

Suppose, for expositionary convenience, that $\bar{w}_{\text{equity}} = 20\%$. How can we introduce life-cycle investing, ie, changes in the optimal bond–equity mix as time passes – into the conceptual framework above? The key is to recognise that investors possess both financial wealth, A, and human capital, H, defined as the present value of future savings.[18] (Slavery having been abolished, human capital is not tradable.) Suppose that an investor, a young investment banker, owns one monetary unit of financial wealth and four units of human capital. Since they own only one unit of assets in which to physically invest (they cannot trade their human capital), they will put 100% of their assets into equities. This translates into 20% equities relative to their total wealth (financial assets plus human capital), assuming that human capital has the character of a bond. The optimal equity allocation, w^*_t, is then

$$w^*_t = \bar{w} + \bar{w}\frac{H_t}{A_t} = 20\% + 20\%\frac{4}{1} = 100\% \tag{2.11}$$

In this equation the first 20% refers to the investor's optimal equity holding if their human capital were zero. The second part reflects the leverage necessary to arrive at 20% equity exposure on their total wealth (ie, five units of capital).

What about the nature of the human capital? Suppose that the investor's job carries risks related to the equity market, so they already own (implicit) equity, say 5%, in their human capital. Would it still be optimal to invest 100% in equities? After taking into account the implicit equity risk, the optimal allocation in equities reduces to

$$w^*_t = \bar{w} + (\bar{w} - \omega_t)\frac{H_t}{A_t} = 20\% + (20\% - 5\%)\frac{4}{1} = 80\% \tag{2.12}$$

The relationship between human and financial capital determines the way assets are allocated over the investor's lifetime. As they grow older, the investor's financial wealth will rise relative to their human capital (we assume that their financial wealth increases with age, which at the same time tends to reduce their human capital). At retirement, their human capital is zero and they own 80% fixed income and 20% equities. With start- and end-points now known, it is interesting to see how the optimal asset allocation evolves over time. Taking two optimal allocations one period apart and subtracting the first from the second, we can write

$$\Delta w^* \approx \frac{H}{A}\Bigg[\bar{w} \times \underbrace{(R_{\mathrm{H}} - R_{\mathrm{p}})}_{\substack{\text{Relative}\\ \text{performance}}} + \underbrace{\Delta\omega}_{\substack{\text{Change in}\\ \text{equity character}}}\Bigg] \tag{2.13}$$

where $R_{\mathrm{H}} - R_{\mathrm{p}}$ denotes the growth differential between human and financial (portfolio) capital. Changes in equity weights are expected to be negative (decreasing optimal equity allocation) as long as asset growth outperforms growth in human capital. This might not always be the case: if the investor is young and job changes lead to fast salary increases, human capital growth might outrun asset growth. This will result in rising equity allocations. However, as the investor approaches retirement, growth in human capital will eventually become negative as the time remaining in terms of professional life shortens. Changes in the equity character of human capital (typically negative, ie, decreasing equity character) will decelerate shifts into fixed income. Rebalancing within financial wealth will be contracyclical (equity will increase if financial wealth underperforms). The most sizeable shifts are likely to occur when investors are young as the ratio of human to financial capital is still large. Assuming no uncertainty in the character of equity, we get, by applying a variance operator to Equation 2.13

$$\sigma^2(\Delta w^*) \approx \left(\frac{H}{A}(\bar{w} - \varpi)\right)^2 \times \sigma^2(R_{\mathrm{H}} - R_{\mathrm{p}}) \tag{2.14}$$

The volatility of changes in the optimal allocation, $\sigma^2(\Delta w)$, is highest if H/A is large, ie, if investors are young. Time does not explicitly enter the equations for lifetime asset allocation, but it implicitly affects asset allocation through changes in human capital. Although the above model ignores everything that makes lifetime asset allocation a difficult problem to solve (we took consumption as exogenous, modelled uncertainty and the character of human capital in a very *ad hoc* way, did not account for elasticity in labour supply, etc), it provides a simple framework that requires little mathematical modelling apart from the calculation of present values and it gives answers that are very similar to much less manageable models.[19] The model shows that even young investors might find it optimal to invest most of their income in bonds if their human capital already exposes them to considerable equity risk. In general, it is not true that investors will always move along the efficient frontier from right to left. The rate at which they make their way along this frontier depends mainly on the relative performance of their human capital and, hence, is highly individual.

APPENDIX A: ONE-STEP VERSUS TWO-STEP OPTIMISATION

When optimising a portfolio one often has to deal with a block structure, in that two or more blocks of assets – eg, equities and bonds, equities and currencies, or active managers and asset classes – must be optimally combined into one portfolio. Very often the correlation between blocks is ignored, or is conveniently set to zero, and the problem is solved either separately – the solution for the first block of assets ignoring that for the second, and vice versa – or in a two-step process in which one first finds the optimal asset allocation and then the optimal currency allocation. We will review the mathematics involved in this type of optimisation problem and illustrate the issues using a currency hedging example.

A2.1 Base model

Optimal currency hedging remains the subject of an ongoing debate between plan sponsors, asset managers and consultants.[20] In the terminology that follows we distinguish between asset returns (local return plus currency return minus domestic cash rate: $a_i = \Delta p_i/p_i + \Delta s_i/s_i - c_h$) and currency returns (local cash rate plus currency movement minus domestic cash rate: $e_i = \Delta s_i/s_i + c_i - c_h$). The covariance matrix of asset and currency returns is assumed to follow the block structure outlined below

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{aa} & \boldsymbol{\Omega}_{ae} \\ \boldsymbol{\Omega}_{ea} & \boldsymbol{\Omega}_{ee} \end{bmatrix} \tag{2.15}$$

where $\boldsymbol{\Omega}_{aa}$ is the $k \times k$ covariance matrix of asset returns expressed in the home currency, $\boldsymbol{\Omega}_{ee}$ is the $k \times k$ covariance matrix of currency returns (assuming that we have as many currencies as assets) and $\boldsymbol{\Omega}_{ae}$ is the $k \times k$ covariance matrix between asset returns and currency returns.[21] Currency hedging takes place in the form of regression hedging, where we regress asset returns against all currency returns to allow for possible cross-correlation

$$a_i = h_i + h_{i1}e_1 + h_{i2}e_2 + \cdots + h_{ik}e_k + \varepsilon_i \tag{2.16}$$

Regression hedging can be expressed in matrix terms as

$$\boldsymbol{h} = \boldsymbol{\Omega}_{ae}\boldsymbol{\Omega}_{ee}^{-1}$$

where the $k \times k$ matrix of regression hedges will contain the regression coefficients of Equation 2.16

$$\boldsymbol{h} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1k} \\ h_{21} & h_{22} & & \\ \vdots & & \ddots & \\ h_{k1} & & & h_{kk} \end{bmatrix} \tag{2.17}$$

We can now define the variance in asset returns that remains unexplained by currency returns

$$\boldsymbol{\Omega}_{a|e} = \boldsymbol{\Omega}_{aa} - \boldsymbol{h}'\boldsymbol{\Omega}_{ee}\boldsymbol{h} \tag{2.18}$$

and rewrite the inverse of the covariance matrix of asset and currency returns as

$$\boldsymbol{\Omega}_{ae}^{-1} = \begin{bmatrix} \boldsymbol{\Omega}_{a|e}^{-1} & -\boldsymbol{\Omega}_{a|e}^{-1}\boldsymbol{h} \\ -\boldsymbol{h}\boldsymbol{\Omega}_{a|e}^{-1} & \boldsymbol{\Omega}_{ee}^{-1} + \boldsymbol{h}\boldsymbol{\Omega}_{a|e}^{-1}\boldsymbol{h}' \end{bmatrix}$$

where we use the results for the inverse of a partitioned matrix

$$\left.\begin{aligned} \begin{bmatrix} \boldsymbol{P}_{11} & \boldsymbol{P}_{12} \\ \boldsymbol{P}_{21} & \boldsymbol{P}_{22} \end{bmatrix}^{-1} &= \begin{bmatrix} \boldsymbol{\Delta}^{-1} & -\boldsymbol{\Delta}^{-1}\boldsymbol{P}_{12}\boldsymbol{P}_{22}^{-1} \\ -\boldsymbol{P}_{22}^{-1}\boldsymbol{P}_{21}\boldsymbol{\Delta}^{-1} & \boldsymbol{P}_{22}^{-1} + \boldsymbol{P}_{22}^{-1}\boldsymbol{P}_{21}\boldsymbol{\Delta}^{-1}\boldsymbol{P}_{12}\boldsymbol{P}_{22}^{-1} \end{bmatrix} \\ \boldsymbol{\Delta} &= \boldsymbol{P}_{11} - \boldsymbol{P}_{12}\boldsymbol{P}_{22}^{-1}\boldsymbol{P}_{21} \end{aligned}\right\} \tag{2.19}$$

We can, for example, check the value of the $\boldsymbol{\Delta}$

$$\begin{aligned} \boldsymbol{P}_{11} - \boldsymbol{P}_{12}\boldsymbol{P}_{22}^{-1}\boldsymbol{P}_{12} &= \boldsymbol{\Omega}_{aa} - \boldsymbol{\Omega}_{ae}\boldsymbol{\Omega}_{ee}^{-1}\boldsymbol{\Omega}_{ae} \\ &= \boldsymbol{\Omega}_{aa} - \boldsymbol{\Omega}_{ae}\boldsymbol{\Omega}_{ee}^{-1}\boldsymbol{\Omega}_{ee}^{-1}\boldsymbol{\Omega}_{ee}^{-1}\boldsymbol{\Omega}_{ae} \\ &= \boldsymbol{\Omega}_{aa} - (\boldsymbol{\Omega}_{ee}^{-1}\boldsymbol{\Omega}_{ae})\boldsymbol{\Omega}_{ee}^{-1}(\boldsymbol{\Omega}_{ee}^{-1}\boldsymbol{\Omega}_{ae}) \\ &= \boldsymbol{\Omega}_{aa} - \boldsymbol{h}\boldsymbol{\Omega}_{ee}^{-1}\boldsymbol{h} \\ &= \boldsymbol{\Omega}_{a|e} \end{aligned}$$

Before we can start to look at the differences between separate and simultaneous optimisation, we have to define the vector of optimal asset and currency weights, $\boldsymbol{w}' = [\boldsymbol{w}_{\mathrm{a}} \quad \boldsymbol{w}_e]$, as well as the vector of expected asset and currency returns, $\boldsymbol{\mu}' = [\boldsymbol{\mu}_a \quad \boldsymbol{\mu}_e]$. We already know from the optimal (simultaneous) solution to an unconstrained currency hedging problem can be written as $\boldsymbol{w}_{\mathrm{sim}}^{*} = \lambda\boldsymbol{\Omega}^{-1}\boldsymbol{\mu}$ (Equation 1.4). However, expanding this expression will generate additional insight.

A2.2 Simultaneous optimisation

We will start with the optimal solution (choosing optimal asset and currency positions simultaneously rather than using a two-step process) to the optimal hedging problem

$$\boldsymbol{w}^*_{\text{sim}} = \begin{bmatrix} \boldsymbol{w}^*_{\text{a,sim}} \\ \boldsymbol{w}^*_{e,\text{sim}} \end{bmatrix} = \begin{bmatrix} \lambda(\boldsymbol{\Omega}^{-1}_{a|e}\boldsymbol{\mu}_a - \boldsymbol{\Omega}^{-1}_{a|e}\boldsymbol{h}'\boldsymbol{\mu}_e) \\ \lambda\boldsymbol{\Omega}^{-1}_{ee}\boldsymbol{\mu}_e - \boldsymbol{h}\boldsymbol{w}^*_{\text{a,sim}} \end{bmatrix} \tag{2.20}$$

The optimal currency position in a simultaneous optimisation involves a speculative as well as a hedging demand

$$\boldsymbol{w}^*_{e,\text{sim}} = \underbrace{\lambda\boldsymbol{\Omega}^{-1}_{ee}\boldsymbol{\mu}_e}_{\substack{\text{Speculative}\\ \text{demand}}} - \underbrace{\boldsymbol{h}\boldsymbol{w}^*_{\text{a,sim}}}_{\substack{\text{Hedge}\\ \text{demand}}} \tag{2.21}$$

If currencies carry a positive risk premium (the currency return is, on average, larger than the interest rate differential), currencies will be included in the optimal portfolio because the first term in Equation 2.21 will be positive. Let us instead focus on the case, often assumed by practitioners, that currencies do not offer a significant risk premium. In this case Equation 2.20 becomes

$$\begin{bmatrix} \boldsymbol{w}^*_{\text{a,sim}} \\ \boldsymbol{w}^*_{e,\text{sim}} \end{bmatrix} = \begin{bmatrix} \lambda\boldsymbol{\Omega}^{-1}_{a|e}\boldsymbol{\mu}_a \\ -\boldsymbol{h}\boldsymbol{w}^*_{\text{a,sim}} \end{bmatrix} \tag{2.22}$$

Suppose that local asset returns are uncorrelated with currency returns. In that case, taking on currency risk does not help to reduce (hedge) asset risk and, as currency risk always comes as an add-on to asset risk, we would intuitively find unitary hedging optimal – currencies only add noise to a portfolio.[22] How does this intuition carry over to Equation 2.22? If local returns, $\Delta p_i/p_i$, for the ith asset are not correlated with currency movements, $\Delta s_i/s_i$, we must find that the covariance between currency returns and foreign asset returns in home currency units contains solely the covariance between currencies[23]

$$\begin{aligned} \operatorname{cov}\left(\frac{\Delta p_i}{p} + \frac{\Delta s_i}{s_i}, \frac{\Delta s_j}{s_j}\right) &= \operatorname{cov}\left(\frac{\Delta p_i}{p} + \frac{\Delta s_i}{s_i}\right) + \operatorname{cov}\left(\frac{\Delta s_i}{s_i}, \frac{\Delta s_j}{s_j}\right) \\ &= \operatorname{cov}\left(\frac{\Delta s_i}{s_i}, \frac{\Delta s_j}{s_j}\right) \end{aligned} \tag{2.23}$$

In matrix terms this becomes

$$\boldsymbol{\Omega}_{ea} = \boldsymbol{\Omega}_{ee} \quad \text{and} \quad \boldsymbol{h} = \boldsymbol{\Omega}_{ea}\boldsymbol{\Omega}^{-1}_{ee} = \mathbf{1}$$

and hence the currency positions will completely hedge out the currency risk that arises from the unhedged asset positions

$$\begin{bmatrix} \boldsymbol{w}^*_{\text{a,sim}} \\ \boldsymbol{w}^*_{e,\text{sim}} \end{bmatrix} = \begin{bmatrix} \lambda \boldsymbol{\Omega}^{-1}_{a|e} \boldsymbol{\mu}_a \\ -\boldsymbol{w}^*_{\text{a,sim}} \end{bmatrix} \tag{2.24}$$

Now suppose the opposite – that foreign asset returns (in home currency units) and currency returns are not correlated. This scenario is very unlikely as, for example, it would be rare for Japanese equity returns in US dollars not to be positively correlated with the US dollar.

Hedging currency risk would now increase total risk as the negative correlation between local returns and currency returns would no longer affect portfolio risk. The optimal hedge ratio becomes zero and positions are given in Equation 2.25

$$\begin{bmatrix} \boldsymbol{w}^*_{\text{a,sim}} \\ \boldsymbol{w}^*_{e,\text{sim}} \end{bmatrix} = \begin{bmatrix} \lambda \boldsymbol{\Omega}^{-1}_{aa} \boldsymbol{\mu}_a \\ 0 \end{bmatrix} \tag{2.25}$$

This follows from $\boldsymbol{\Omega}_{ea} = \mathbf{0}$, so $\boldsymbol{h} = \boldsymbol{\Omega}_{ea}\boldsymbol{\Omega}^{-1}_{ee} = \mathbf{0}$ and, hence, the conditional asset volatility will equal the unconditional volatility: $\boldsymbol{\Omega}_{a|e} = \boldsymbol{\Omega}_{aa} - \boldsymbol{h}'\boldsymbol{\Omega}_{ee}\boldsymbol{h} = \boldsymbol{\Omega}_{aa}$, where $\boldsymbol{\Omega}_{a|e}$ denotes the covariance of asset returns conditional on currency returns.

We can summarise these findings into a quick taxonomy of currency hedging based on the behaviour of local returns and currency returns. If currencies carry a risk premium, there will always be a speculative aspect to currency exposure. However, if we assume a zero risk premium to currencies, we have to review currency exposure in terms of its ability to reduce asset risk. Zero correlation between local returns and currency returns will make currencies a noisy asset that adds risk without providing compensation in the form of returns or diversification. Negative correlation between local returns and currency returns makes currencies a hedge asset that reduces total portfolio risk. Hedging out currency risk completely would increase total portfolio risk. Positive correlation between local returns and currency returns would achieve the opposite. In that case over-hedging (where the short position in currencies is larger than the long position in assets carrying currency risk) is optimal.

A2.3 Two-step optimisation

So far we have looked at the simultaneous optimisation of currency and asset positions. Alternatively, we could optimise asset positions in a first step and in the second step choose optimal currency positions conditional on the already established asset positions. The result of this approach, called "partial optimisation", is given below

$$\boldsymbol{w}^*_{\text{par}} = \begin{bmatrix} \boldsymbol{w}^*_{\text{a,par}} \\ \boldsymbol{w}^*_{e,\text{par}} \end{bmatrix} = \begin{bmatrix} \lambda \boldsymbol{\Omega}^{-1}_{aa} \boldsymbol{\mu}_a \\ \lambda \boldsymbol{\Omega}^{-1}_{ee} \boldsymbol{\mu}_e - \boldsymbol{h} \boldsymbol{w}^*_{\text{a,par}} \end{bmatrix} \tag{2.26}$$

Terms representing conditional covariance drop out (as by definition it is not taken into account) and there is no feedback of currency positions on asset positions as this has previously been ignored.

Clearly, partial optimisation is sub-optimal (leads to a lower utility) as it ignores the covariances between assets and currencies.

A2.4 One-step (separate) optimisation

The final option for constructing portfolios in the presence of currencies is separate optimisation, also known as "currency overlay"

$$\boldsymbol{w}^*_{\text{sep}} = \begin{bmatrix} \boldsymbol{w}^*_{\text{a,sep}} \\ \boldsymbol{w}^*_{e,\text{sep}} \end{bmatrix} = \begin{bmatrix} \lambda \boldsymbol{\Omega}^{-1}_{aa} \boldsymbol{\mu}_a \\ \lambda \boldsymbol{\Omega}^{-1}_{ee} \boldsymbol{\mu}_e \end{bmatrix} \tag{2.27}$$

"Separate" refers to the fact that asset and currency decisions are taken by different departments in an investment firm which do not take into account each other's positions in a client portfolio. Separate optimisation is dominated by partial optimisation. This can be seen by calculating the difference in utility, U, that both weightings would create

$$\begin{aligned} U(\boldsymbol{w}^*_{\text{par}}) - U(\boldsymbol{w}^*_{\text{sep}}) & \\ &= (\boldsymbol{w}^*_{\text{par}} - \boldsymbol{w}^*_{\text{sep}})\boldsymbol{\mu} - \frac{1}{2\lambda}(\boldsymbol{w}^{*\prime}_{\text{par}} \boldsymbol{\Omega} \boldsymbol{w}^*_{\text{par}} - \boldsymbol{w}^{*\prime}_{\text{sep}} \boldsymbol{\Omega} \boldsymbol{w}^*_{\text{sep}}) \\ &= \tfrac{1}{2}\lambda(\boldsymbol{\mu}'_a \boldsymbol{\Omega}^{-1}_{aa} \boldsymbol{\Omega}_{ae} \boldsymbol{\Omega}^{-1}_{ee} \boldsymbol{\Omega}_{ea} \boldsymbol{\Omega}^{-1}_{aa} \boldsymbol{\mu}_a) \end{aligned} \tag{2.28}$$

As long as $\boldsymbol{\Omega}^{-1}_{aa} \boldsymbol{\Omega}_{ae} \boldsymbol{\Omega}^{-1}_{ee} \boldsymbol{\Omega}_{ea} \boldsymbol{\Omega}^{-1}_{aa}$ is positive-definite this will always be the case. Partial and separate optimisation yield the same result if investors find zero hedging optimal.

A2.5 Inclusion of constraints

To conclude this appendix, some constraints will be added to the above solution. As stated in Section 1.1, the optimal solution to the constrained optimisation problem is

$$\begin{aligned} \boldsymbol{w}^*_{\text{sim}} &= \begin{bmatrix} \boldsymbol{w}^*_{\text{a,sim}} \\ \boldsymbol{w}^*_{e,\text{sim}} \end{bmatrix} \\ &= \boldsymbol{\Omega}^{-1}\boldsymbol{A}'(\boldsymbol{A}\boldsymbol{\Omega}^{-1}\boldsymbol{A}')^{-1}\boldsymbol{b} + \lambda\boldsymbol{\Omega}^{-1}(\boldsymbol{I} - \boldsymbol{A}'(\boldsymbol{A}\boldsymbol{\Omega}^{-1}\boldsymbol{A}')\boldsymbol{A}\boldsymbol{\Omega}^{-1})\boldsymbol{\mu} \end{aligned} \tag{2.29}$$

If we define

$$\left.\begin{aligned} \boldsymbol{A}'(\boldsymbol{A}\boldsymbol{\Omega}^{-1}\boldsymbol{A}')^{-1}\boldsymbol{b} &= \begin{bmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \end{bmatrix} \\ \boldsymbol{I} - \boldsymbol{A}'(\boldsymbol{A}\boldsymbol{\Omega}^{-1}\boldsymbol{A}')^{-1}\boldsymbol{A}\boldsymbol{\Omega}^{-1} &= \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \end{aligned}\right\} \tag{2.30}$$

we can expand the optimal solution to

$$\begin{aligned} &\begin{bmatrix} \boldsymbol{w}^*_{\text{a,sim}} \\ \boldsymbol{w}^*_{e,\text{sim}} \end{bmatrix} \\ &\quad= \begin{bmatrix} \boldsymbol{\Omega}^{-1}_{a|e}\boldsymbol{B}_1 - \boldsymbol{\Omega}^{-1}_{a|e}\boldsymbol{h}'\boldsymbol{B}_2 \\ \boldsymbol{\Omega}^{-1}_{ee}\boldsymbol{B}_2 - \boldsymbol{h}(\boldsymbol{\Omega}^{-1}_{a|e}\boldsymbol{B}_1 - \boldsymbol{\Omega}^{-1}_{a|e}\boldsymbol{h}'\boldsymbol{B}_2) \end{bmatrix} \\ &\qquad+ \lambda \begin{bmatrix} \boldsymbol{\Omega}^{-1}_{a|e}(\boldsymbol{A}_{11}\boldsymbol{\mu}_a + \boldsymbol{A}_{12}\boldsymbol{\mu}_e - \boldsymbol{h}'\boldsymbol{A}_{21}\boldsymbol{\mu}_a - \boldsymbol{h}'\boldsymbol{A}_{22}\boldsymbol{\mu}_e) \\ \boldsymbol{\Omega}^{-1}_{ee}\boldsymbol{A}_{21}\boldsymbol{\mu}_a + \boldsymbol{\Omega}^{-1}_{ee}\boldsymbol{A}_{22}\boldsymbol{\mu}_e \\ -\boldsymbol{h}\boldsymbol{\Omega}^{-1}_{a|e}(\boldsymbol{A}_{11}\boldsymbol{\mu}_a + \boldsymbol{A}_{12}\boldsymbol{\mu}_e - \boldsymbol{h}'\boldsymbol{A}_{21}\boldsymbol{\mu}_a - \boldsymbol{h}'\boldsymbol{A}_{22}\boldsymbol{\mu}_e) \end{bmatrix} \end{aligned} \tag{2.31}$$

Although Equation 2.31 looks daunting, it offers a closed-form solution for the optimal hedging problem. It could be used to visually inspect the loss in efficiency when partial rather than simultaneous optimisation is used. However, constraints will in general limit the loss of efficiency – the full correlation structure is limited by the set of constraints and therefore cannot be taken advantage of fully.

[1] The selection of asset classes suitable for optimisation is both quantitative and judgmental.

[2] See Appendix A (on page 66).

[3] A recommended introduction to the practical implementation of cluster analysis is Kaufman and Rousseeuw (1990).

[4] This method is also called “agglomerative”, whereas methods that begin by uniting all assets in one cluster are termed “divisive”.

[5] This simple distance function satisfies all required properties for a distance function: the distance to itself is zero; the distance between A and B is the same as that between B and A (symmetry); and the straight-line distance between A and C is always shorter than when a detour is made via B.

[6] Historical data sets are available from Hedge Fund Research’s website at URL: http://www.hfr.com.

7 Insightful's S-PLUS contains routines for alternative clustering methods.

8 As a consequence, corporate high yield tends to dominate all other fixed-income classes, leaving little room for emerging market debt (typically about 17% volatility per annum for JP Morgan EMBI) at almost all risk levels.

9 See Blundell and Ward (1987).

10 Note that this procedure leaves the average return virtually unchanged (for a large enough number of observations, T)

$$\bar{r}_t^* = \frac{1}{1-a_1}\bar{r} - \frac{a_1}{1-a_1}\bar{r} = \bar{r}$$

but increases the variance estimate. Applying the variance operator to Equation 2.5, we arrive at

$$\sigma^2(r^*) = \frac{1+a_1^2}{(1-a_1)^2}\sigma^2 \operatorname{var}(r)$$

UK property volatility rises to about 8% (instead of the naive 3% estimate) and high-yield volatility climbs to 9% (from 6%). Both figures appear much more reasonable than the widely used naive estimates.

11 These sub-indexes are not representative of the whole hedge fund universe as the aim here is merely to illustrate the methodology.

12 See Scholes and Williams (1977).

13 This statement has been written from a total risk view. If investors look at benchmark-relative risks (or, equivalently, to long–short portfolios), an increase in correlation means that the hedge becomes better.

14 This concept is taken from Chow *et al* (1999).

15 This distinguishes the analysis presented in this chapter from approaches that focus on measuring downside correlation by defining the down market with respect to a core asset.

16 One caveat of the described methodology, though, is that Table 2.3 might change if new assets are included. This results directly from the fact that outliers are defined by looking at all assets simultaneously. Thus, what has previously been a normal observation without the new asset could become unusual if the included asset shows an unusual movement from this data point.

17 It can be shown that, for an investor with utility $u(1+R) = (1-\gamma)^{-1}(1+R)^{1-\gamma}$ the optimal solution is given by

$$w_{\text{equity}} = \frac{1}{\gamma}\frac{\mu}{\sigma^2}$$

However, if we introduce the probability of dying in year t as δt, this changes to

$$w_{\text{equity}} = \frac{1}{\gamma}\frac{\mu - \delta t}{\sigma^2}$$

One direct consequence of this is that women hold more equities than men as a woman is less likely to die at age t than a man.

18 This model follows Scherer and Ebertz (1998).

19 See Campbell and Viceira (2002, Chapters 7 and 8).

20 This section follows the work by Jorion (1994).

21 For example, element 1,2 in $\boldsymbol{\Omega}_{aa}$ is calculated as

$$\operatorname{cov}(a_1, a_2) = \operatorname{cov}\left(\frac{\Delta p_1}{p_1} + \frac{\Delta s_1}{s_1} - c_h, \frac{\Delta p_2}{p_2} + \frac{\Delta s_2}{s_2} - c_h\right)$$

22 Unitary hedging: currency risk that comes automatically with asset risk is hedged on a one-to-one basis; for example, a 10% position in US assets is hedged with an offsetting −10% position in the US dollar.

23 We suppress cash rates as they are known in advance and do not change the risk numbers.

REFERENCES

Blundell, G., and C. Ward, 1987, "Property Portfolio Allocation: A Multifactor Model", *Land Development Studies* 4, pp. 145–56.

Campbell, J., and L. Viceira, 2002, *Strategic Asset Allocation: Portfolio Choice for Long-Term Investors* (Oxford University Press).

Chow, G., E. Jacquier, M. Kritzman and K. Lowry, 1999, "Optimal Portfolios in Good Times and Bad", *Financial Analysts Journal* 55, pp. 65–73.

Connor, G., L. R. Goldberg and R. A. Korajceck, (2009), *Portfolio Risk Management* (Princeton, NJ: Princeton University Press).

Jorion, P., 1994, "Mean Variance Analysis of Currency Overlays", *Financial Analysts Journal* 50, pp. 48–56.

Kaufman, L., and P. Rousseeuw, 1990, *Finding Groups in Data: An Introduction to Cluster Analysis* (New York: John Wiley & Sons).

Scherer, B., and T. Ebertz, 1998, "A Simple Model for Lifetime Asset Allocation", *Journal of Private Portfolio Management* 1, pp. 27–30.

Scholes, M., and J. Williams, 1977, "Estimating Betas from Nonsynchronous Data", *Journal of Financial Economics* 14, pp. 327–48.

3

Incorporating Deviations from Normality: Lower Partial Moments

This chapter deals with non-normality, a prominent shortcoming of traditional portfolio analysis. We first review key issues that arise when we are faced with non-normality in data series. The main focus of the chapter, however, is the application of lower partial moments as one way of dealing with asymmetric return distributions. A second, more general method will be presented in Chapter 5.

3.1 NON-NORMALITY IN RETURN DATA

3.1.1 Single-period returns: visualising and testing for non-normality

This and the next two sections will deal with non-normality (which was identified in Chapter 1 as a potential shortcoming of the traditional Markowitz framework) and its impact on portfolio choice. We will not attempt to arrive at some definitive "cookbook recipe" for constructing a portfolio but, rather, to establish the key issues to keep in mind when doing so. These are:

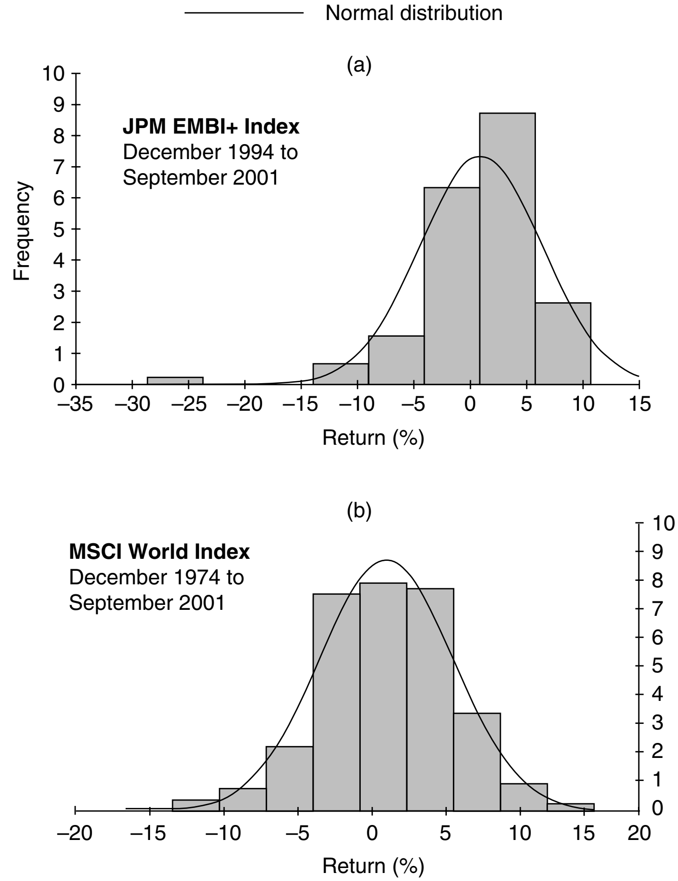
- Are returns normal?
- Are deviations from normality statistically significant?
- Are these deviations stable, ie, forecastable?
- Will non-normality vanish over time?
- Can we model a simple non-normal alternative?

Most of these questions are covered in this section, though the last two are answered in Sections 3.1.2 and 3.1.3, respectively.

3.1.1.1 Returns are not normal

The first question is easy enough to answer: as is recognised in asset management, returns are not normally distributed.[1] This is

Figure 3.1 Departure from normality of different indexes: comparison of emerging market bonds (a) with world equities (b) and normal distribution



important because the assumption of normality makes calculations involving risk much simpler than they are with non-normal alternatives, allowing one to easily aggregate asset risks both over time and across assets.[2] Most financial models rely heavily on the normality assumption.[3]

However, the degree of non-normality differs between financial time series. This can be seen by comparing histograms (empirical distributions) of returns for world equities and emerging market bonds with a normal distribution. This is done using the MSCI World Index and the JP Morgan EMBI+ Index in Figure 3.1.

Returns on equities in the MSCI World Index, which represents many stocks diversified across regions and sectors, are closer to normality than returns on emerging market bonds, where the JP Morgan EMBI+ Index contains a small number of bonds with a regional concentration in Latin America. This is hardly surprising, as – provided that they show little correlation – the more assets an index contains, the smaller the deviation from non-normality will be. However, if correlation increases at certain times (as seen recently in emerging markets), extreme returns in single holdings will not be diversified away and the empirical distribution will show significant negative "skewness".[4] Ignoring non-normality in developed markets presents less of a problem than it does in emerging markets.[5] Hence, it is necessary to judge each type of market on its merits as to whether it is worth the effort to model non-normality explicitly.

3.1.1.2 Statistical significance of deviations

If we obtain an empirical data sample like a financial time series we would not, of course, expect a relatively small number of random drawings from it to be distributed exactly normally. However, we need some way other than visual inspection of measuring non-normality so that we have an objective basis for judging whether deviations from normality are statistically significant.

Deviations from normality are often summarised in two measures known as "skewness" and "kurtosis". Skewness measures the asymmetry of a distribution; a distribution is negatively skewed if it has a long tail to the left and positively skewed if its long tail is to the right. Skewness can be calculated (and tested for significance) by using

$$\text{Skew} = \frac{1}{T}\sum_{i=1}^{T}\frac{(r_i-\mu)^3}{\hat{\sigma}^3} \sim N\left(0, \sqrt{\frac{6}{T}}\right) \tag{3.1}$$

where T is the number of observations, r_i is the return in period i, μ is the mean, σ is the standard deviation, and N stands for normal distribution. A symmetric distribution (like the normal bell curve) has zero skewness as all realisations are equally scattered to the left and right of the mean.

Kurtosis measures the "peakedness" of a distribution, ie, the size of the peak relative to the tails. The distributions of illiquid asset classes or those with periodic jumps are flatter than the normal distribution as they are characterised by a high probability of small movements and large jumps (both positive and negative), along with a low probability of intermediate movements. These characteristics give their distributions "fat" tails, ie, values tend to spread into the extremes of the distribution. Excessive kurtosis (relative to the kurtosis of three for a normal distribution) can be calculated, and tested for significance, using

$$\text{Kurtosis} = \frac{1}{T}\sum_{i=1}^{T}\frac{(r_i-\mu)^4}{\hat{\sigma}^4} - 3 \sim N\left(0, \sqrt{\frac{24}{T}}\right) \tag{3.2}$$

We can use both measures *and* simultaneously test for normality using the Jarque–Bera (JB) test[6]

$$\text{JB} = T(\tfrac{1}{6}\text{Skewness}^2 + \tfrac{1}{24}\text{Kurtosis}^2) \sim \chi^2(2) \tag{3.3}$$

which has a chi-squared distribution with two degrees of freedom.[7] Applying the JB test to real data gives the results in Table 3.1. With the exception of US government bonds, almost all series show statistically significant deviations from normality at the 5% level, indicated by a p value below 0.05. The results also show that even an index with relatively few and highly correlated assets, such as the MSCI EMF, can be almost normally distributed, contrary to the rule of thumb mentioned earlier.

3.1.1.3 Stability of deviations: "persistency"

Statistically significant non-normality alone is generally not enough to justify the effort to model non-normality. To make them exploitable deviations from normality have to be stable, or persistent, across time periods. Otherwise, the error maximisation property in the portfolio optimisation process takes effect and portfolios that are deliberately constructed to show positive in-sample skewness will fail out-of-sample.

Unfortunately, authors are divided on the persistency issue.[8] Rather than reviewing the literature, we offer a test for persistency of skewness that can easily be applied to determine whether it is worth modelling non-normality for particular assets:[9]

1. Choose an asset universe – emerging markets, for example.
2. Split the data set into two non-overlapping time periods of equal length.

Table 3.1 Testing for normality: application of Jarque–Bera test to selected indexes (monthly data)

	JPM US Gov.	MSCI EMF	SB WGBI	JPM EMBI+	MSCI World
Mean (%)	0.69	−0.57	0.73	0.92	0.98
Maximum (%)	4.35	16.71	4.02	10.71	14.99
Minimum (%)	−2.69	−29.22	−2.01	−28.74	−16.63
Volatility (%)	1.31	8.55	1.05	5.45	4.55
Skewness	0.05	−0.64	0.12	−1.97	−0.22
Kurtosis	3.11	3.96	3.86	11.22	3.91
JB test	30.17	6.12	6.65	21.62	13.62
p value	0.92	0.05	0.04	0.00	0.00
No. observations	189	57	201	93	321

Data for all series end September 2001 – longest series available.
JPM US Gov., JPM US Government Bonds; MSCI EMF, Morgan Stanley Capital International Emerging Markets Free Index; SB WGBI, Salomon Brothers World Government Bond Index; JPM EMBI+, JP Morgan Emerging Markets Bond Index Plus; MSCI World, MSCI Emerging Markets Free.

3. Calculate skewness, kurtosis or asymmetric semivariance (given by the semivariance minus half the variance) for each time period.[10]
4. Run a regression of the form

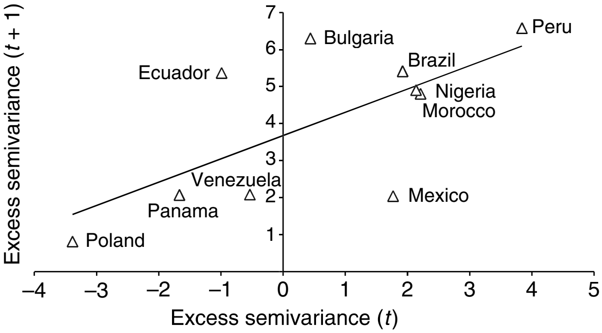
$$\text{skew} - \text{measure}_{t+1} = a + b \cdot \text{skew} - \text{measure}_t + \varepsilon_t \tag{3.4}$$

   Persistency is indicated by a significant slope parameter and a high R^2.

This test will be illustrated with data for emerging debt markets – specifically, data from JP Morgan EMBI+ sub-indexes for the countries included in Figure 3.2.

The data runs from August 1994 to November 2001 and is split into time series of equal length. We run the regression given by Equation 3.4 and use the excess semivariance as a measure of skewness. The resulting fit is also included in Figure 3.2. The t value on the slope coefficient (0.63) is statistically significant (2.53), while the regression explains 44% of the variance in excess skewness. The result indicates that there is persistency in the excess skewness and justifies – indeed calls for – the modelling of non-normality in these emerging market debt returns.

Figure 3.2 Test for persistence of excess skewness in data from JP Morgan EMBI+ country sub-indexes



3.1.2 Normality and multi-period returns

When defending the assumption that multi-period returns are distributed normally we have a powerful ally: the Central Limit Theorem (CLT). This states that the sum of independent and identically distributed variables converges to a normal distribution as long as the variables have finite variances.[11] Convergence takes place after approximately 30 random numbers have been added up. Correspondingly, the product of random variables gives *lognormally* distributed variables.[12] Hence, multi-period returns are better described using a lognormal distribution as they are effectively products of single-period returns.

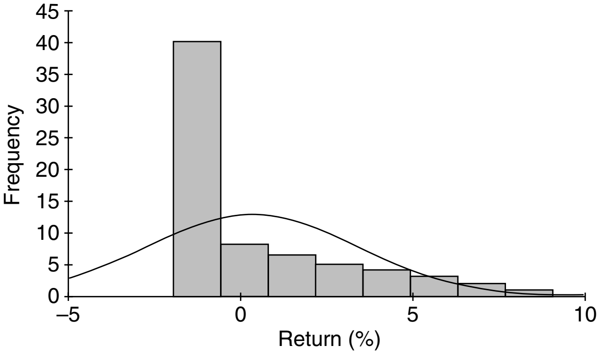
3.1.2.1 Convergence to normality: simulation example

We can simulate the consequences of the CLT with one of the most non-normally distributed returns encountered in portfolio construction exercises: portfolio insurance.[13] Suppose that we engage in a monthly protective put strategy by buying one at-the-money put option per asset. We will assume that the underlying asset returns 0.83% per month with a volatility of 5%. The put option price is about 2% (assuming a 3.5% annual risk-free rate). Monthly returns for this strategy can be calculated as

$$R_{\text{month},i} = \frac{\max[100 - S_i, 0] + S_i}{100 + \text{Put}} = \frac{\max[100, S_i]}{100 + \text{Put}} \tag{3.5}$$

where the asset price is normalised to 100 at the beginning of each month. The histogram of 10,000 monthly draws and a fitted normal distribution are shown in Figure 3.3.

Figure 3.3 Portfolio insurance example: histogram of 10,000 simulated monthly protective put returns with fitted normal distribution



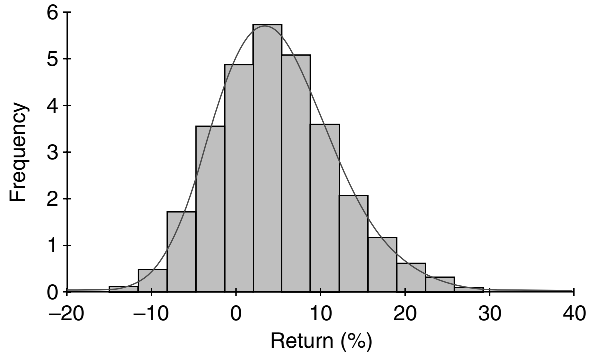
Portfolio insurance transforms small probabilities of large losses into large probabilities of small losses. Volatility now becomes a seriously misleading concept as it massively overestimates the probability of losses. In fact, it becomes more a measure of upside potential than of downside risk. Negative one-sigma events cannot occur as the put protection compensates for losses in the underlying asset. The positive skewness in the distribution makes the average return an easily misinterpreted number. This is sometimes deliberately used in marketing pitches as, for skewed distributions, mean and median returns can be very different: distributions with a positive skew have a mean return that is much higher than the median return. The average return is no longer the outcome where about half the returns are lower and half are higher. In fact, in this example, returns will fall below the average in about 65% of all cases (with a negative median). Thus, measuring the costs of portfolio insurance by the difference in mean returns is potentially misleading.

Suppose now that we continue this strategy over 30 months, ie, using 30 monthly resets of monthly protective puts. The annualised return in this case is given by

$$R_{30\text{ months}} = \frac{30}{12}\sqrt{\prod_{i=1}^{30}(1 + R_{\text{month},i})}$$

The results derived from 10,000 simulations of rolling over 30 monthly protective puts can be seen in Figure 3.4. After 30 periods our very non-normal protective put returns have converged, as expected, to

Figure 3.4 Histogram of 30 monthly protective put returns (simulated)



a lognormal distribution (the solid curve). This is why mainstream asset allocators feel little need to model non-normality in long-term asset allocation studies.[14]

3.1.2.2 Convergence to normality: some problems with real data

So far, this chapter has considered a textbook case. The Monte Carlo simulations in the previous section were drawn from a normal distribution of asset returns with finite variance, and protective put returns were generated using Equation 3.5 under the assumption (construction) that asset returns are drawn independently.[15] Doing so leads to the familiar convergence to normality, as seen in the CLT. With real data, however, the assumptions that allow us to invoke the CLT might not be realistic.[16]

First, although single-period returns may be uncorrelated, they are not independent, as the CLT requires. For example, volatility tends to cluster, in that high- and low-volatility regimes tend to persist for a while; squared returns (a proxy for volatility) are significantly correlated; and, in general, Garch models – time series models that capture dependency in second moments – fit the data reasonably well.[17]

A second, more technical, argument is that return distributions do not necessarily exhibit finite variance. In fact, if single-period returns show infinite variance (distributions with a tail index $\alpha \leqslant 2$), the same logic that underlies the CLT also provokes convergence to a class of fat-tailed distributions. Their sum will then converge to a so-called "α-stable distribution" (which includes the normal distribution as a special case if $\alpha = 2$). The parameter α plays an important role in Extreme Value Theory. Estimation and inference of α is relatively straightforward.[18] It has been shown that, if we order returns $r_1 \leqslant r_2 \leqslant \cdots \leqslant r_{T_{\text{tail}}} \leqslant \cdots \leqslant r_T$, we can estimate the tail index from

$$\hat{\alpha}^{-1} = \frac{1}{T_{\text{tail}}} \sum_{i=T_{\text{tail}}}^{1} \log\left(\frac{r_i}{r_{T_{\text{tail}}+1}}\right) \tag{3.6}$$

where T_{tail} is the number of observations in the tail and $T_{\text{tail}} + 1$ is the observation where the tail is assumed to start. Unfortunately, there is no hard and fast rule on how to determine T_{tail}. Practitioners plot the tail index against the choice of T_{tail} – in what is known as a "hill" plot – and choose the tail index where the curve becomes more or less horizontal (the tail index is robust against changes in the number of tail observations).[19] To test the null hypothesis, H_0, that $\alpha < 2$ against $H_1 : \alpha \geqslant 2$ (the value of two divides distributions with fat tails but finite variance from distributions with infinite variance), we can use[20]

$$\sqrt{T_{\text{tail}}}(\hat{\alpha} - \alpha) \sim N(0, \alpha^2) \tag{3.7}$$

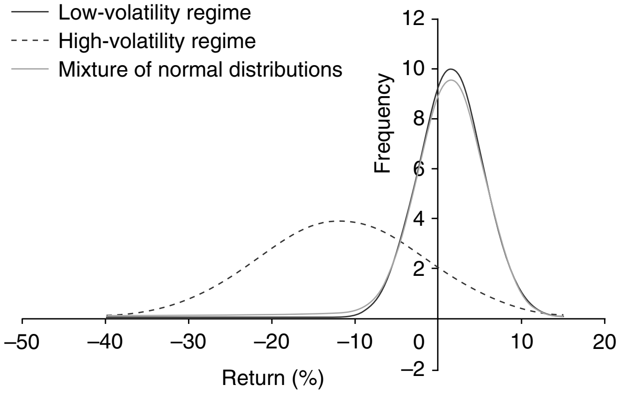
It should be noted that, to be reliable, this test procedure needs extreme values to be present in the data set (as is generally typical of Extreme Value Theory); hence we need enough data to be reasonably sure that extreme events have been sampled in the empirical distribution.[21]

3.1.3 Modelling non-normality with a mixture of normal distributions

Models that can be understood intuitively and fit the data well have the greatest chance of receiving attention from practitioners. As market participants tend to think in terms of regimes – eg, periods of different volatility – it is natural to model a distribution as a combination of two (or more; see Hamilton 1994) normal distributions, with each distribution representing a different regime. Shifts from one regime to another take place randomly with a given probability. Moreover, mixtures of normal distributions can model skewness as well as kurtosis and generally fit the data much better than a non-normal alternative.[22]

The probability density function for a mixture of two normal distributions is a probability-weighted sum of normal density

Figure 3.5 Use of a mixture of two normal distributions to represent JPM EMBI+ data, December 1994 to September 2001



functions[23]

$$
\left.\begin{aligned}
f_{\text{MoN}}(r) &= pf_{\text{high}}(r) + (1-p)f_{\text{low}}(r) \\
f_{\text{high}}(r) &= \frac{1}{\sqrt{2\pi\sigma_{\text{high}}}} \exp\left(-\frac{1}{2}\frac{(r-\mu_{\text{high}})^2}{\sigma^2_{\text{high}}}\right) \\
f_{\text{low}}(r) &= \frac{1}{\sqrt{2\pi\sigma_{\text{low}}}} \exp\left(-\frac{1}{2}\frac{(r-\mu_{\text{low}})^2}{\sigma^2_{\text{low}}}\right)
\end{aligned}\right\} \tag{3.8}
$$

where $f_{\text{MoN}}(r)$ is the combined density of the high-volatility, $f_{\text{high}}(r)$, and low-volatility, $f_{\text{low}}(r)$, regimes, while p denotes the probability of experiencing a draw from the high-volatility regime. The log-likelihood function, $\log(L)$, can be written in the usual way

$$
\begin{aligned}
\log(L) &= \log\left(\prod_{i=1}^{T} f_{\text{MoN}}(r_i)\right) = \sum_{i=1}^{T} \log(f_{\text{MoN}}(r_i)) \\
&= \sum_{i=1}^{T} \log\left(p\frac{1}{\sqrt{2\pi\sigma_{\text{high}}}} \exp\left(-\frac{1}{2}\frac{(r-\mu_{\text{high}})^2}{\sigma^2_{\text{high}}}\right)\right. \\
&\qquad\qquad \left. + (1-p)\frac{1}{\sqrt{2\pi\sigma_{\text{low}}}} \exp\left(-\frac{1}{2}\frac{(r-\mu_{\text{low}})^2}{\sigma^2_{\text{low}}}\right)\right)
\end{aligned} \tag{3.9}
$$

We can now apply maximum-likelihood estimation by maximising Equation 3.9 with respect to μ_{high}, σ^2_{high}, μ_{low}, σ^2_{low} and p. If applied to the EMBI+ data used in Section 3.1, we get the frequency distribution represented by the grey curve in Figure 3.5.

Essentially, the EMBI+ data can be split into a high-volatility (10.22%) and a low-volatility (3.99%) regime. High volatility arises from market jumps, which are, on average, negative (−11.6%). High-volatility regimes have a modest probability (5.3%) and, hence, a limited influence on the combined distribution (which is found by probability-weighting both regimes and adding them up for every return level).

Figure 3.5 shows that if there are two regimes and one has a much greater probability of occurrence than the other, fitting a mixture of two normal distributions does not necessarily result in a bimodal distribution. However, a mixture is preferable as it allows for distributions that are skewed to the left, thus capturing the long left tail in the data much better than the normal alternative.[24]

3.2 LOWER PARTIAL MOMENTS

3.2.1 Illustration of approach using empirical distribution and single return series

So far this chapter has established that, depending on the nature of the data and the time horizon, the assumption of normality might not be close enough to reality. Now we want to describe how non-normality can be incorporated in portfolio construction applications.

We start with the general observation that uncertain returns, $\tilde{R}$, can be decomposed into a threshold return, γ, plus an upside measure, expressed by $\max[\tilde{R} - \gamma, 0]$, which is either positive or zero, minus a downside measure, denoted by $\max[\gamma - \tilde{R}, 0]$, which is also either positive or zero (all measures are relative to the threshold). In combination we get

$$\tilde{R} = \underbrace{\gamma}_{\text{threshold}} + \underbrace{\max[\tilde{R} - \gamma, 0]}_{\text{upside}} - \underbrace{\max[\gamma - \tilde{R}, 0]}_{\text{downside}} \tag{3.10}$$

Suppose that our threshold return is −5%. A return of −15% can then be expressed as

$$-5\% + \underbrace{\max[-15\% - (-5\%), 0]}_{0\%} - \underbrace{\max[-5\% - (-15\%), 0]}_{10\%} = -15\%$$

Measures that try to capture the downside of a return distribution are called lower partial moments, while measures that focus on the upside of a distribution are called upper partial moments. The value of risk measures that capture non-normality increases the more

Table 3.2 Choice of threshold return

Threshold return (γ)	Objective
Zero return	Nominal capital protection
Inflation	Real capital protection
Risk-free rate	Target minimum-opportunity costs
Actuarial rate	Actuarial funding protection
Moving target (benchmark)	Target opportunity costs

return distributions deviate from normality. Moreover, as investors seem to be primarily concerned about downside deviations, some authors call for more behaviourally motivated risk measures that recognise the gap between theoretical risk measures and those used by practitioners.[25]

Lower partial moments are characterised by a threshold – if returns fall beneath this threshold, a "risky" scenario is indicated – and large negative deviations from the threshold are penalised. Starting from Equation 3.10, the lower partial moment of the *m*th degree (which defines the penalty) and threshold, γ, in its *ex post* form can be written as[26]

$$\left.\begin{aligned} \mathrm{lpm}(\gamma, m) &= E\max[(\tilde{R}-\gamma)^m, 0] = \frac{1}{T}\sum_{i=1}^{T}(\tilde{R}_i-\gamma)^m d_i \\ d_i &= \begin{cases} 0, & R_i > \gamma \\ 1, & R_i \leqslant \gamma \end{cases} \end{aligned}\right\} \qquad (3.11)$$

where the function d_i (indicator function) counts the cases in which the *ex post* return is at the threshold value. The parameter *m* describes how the penalty function is shaped.

What can we use to determine our choice of threshold return and the curvature of the penalty function in Equation 3.11? Popular threshold values are based on actuarial target returns (in the case of pension funds), loss aversion (ie, a target return of zero), or purchasing power protection (where the threshold return equals average period inflation). These and other economically plausible possibilities are summarised in Table 3.2.

Effectively, the indicator function decides which observations will enter the calculations in Equation 3.11. If the threshold is set very much to the left of a distribution, relatively few observations will determine the calculation of the respective lower partial moment as extreme events are rare by definition.

The second parameter we control is m. When m is equal to zero, we get the shortfall probability – often used in institutional asset management – denoted by $\mathrm{lpm}(\gamma, 0)$. Anything to the power of zero equals one, and hence Equation 3.11 will add up any value of one for which the indicator function is itself one. Dividing by the total number of observations, we get the probability of underperformance. A key problem with the $\mathrm{lpm}(\gamma, m)$ risk measure is that all underperformance is perceived to be equally undesirable; thus a threshold underperformance by five basis points would have the same impact on our risk measure as an underperformance by 2,500bp. Most practitioners feel uncomfortable with $\mathrm{lpm}(\gamma, 0)$ as this might actually lead to risk-seeking behaviour if the return requirement is so high that only the most lucky draw from the most risky asset could help. Setting $m = 1$, we get $\mathrm{lpm}(\gamma, 1)$, which gives us the average underperformance.[27] As $\mathrm{lpm}(\gamma, 1)$ implies risk-neutrality for all return realisations below the threshold (only the average counts, not the dispersion), it also determines the boundary between risk-averting and risk-seeking behaviour. Although there is no theoretical limit to increase in m, for portfolio construction exercises we will restrict ourselves to $\mathrm{lpm}(\gamma, 2)$, which can be interpreted as the shortfall variance.[28] This measures the dispersion of underperformance and comes closest in interpretation to the usual variance measure.[29] The higher m is, the more weight is given to underperformance for a given distribution. The more skewed a distribution, the higher the value for a given lower partial moment. This can be illustrated as follows.

Suppose we have two assets with the probability distributions given in Table 3.3 and that we choose a threshold return equal to the distribution mean, μ. Although both assets show the same mean and variance, asset two provides much less downside risk than asset one as its distribution is less skewed. This is clearly shown by the risk measure results in the table, the lower partial moments for $m \geqslant 2$ attributing less risk to asset two. However, the calculations also show that the ranking of the two assets depends on the choice of m; for example, asset one looks favourable if we consider shortfall risk but inferior when we look at shortfall variance. The choice is therefore one investors should consider carefully.

Table 3.3 Application of lpm risk measure to two hypothetical assets

	Asset 1		**Asset 2**	
Return (%)	−0.82	21.61	10.39	32.82
Probability (%)	25	75	75	25
Moments of distribution				
Mean, μ	16		16	
Variance, σ	9.71		9.71	
Skew	−1.15		0.04	
Risk measure **(lpm(γ,m))**				
lpm: 16, 0	25%		75%	
lpm: 16, 1	−4.21		−4.21	
lpm: 16, 2	70.75		23.58	
lpm: 16, 3	−1190		−132.2	

First two rows give downside and upside performance for the two assets.

3.2.1.1 Calculation of Equation 3.11: an example

A simple example will help to understand the mechanics of calculating Equation 3.11 when real data is involved. Suppose we obtain the return realisations given in Table 3.4 when we use a threshold return of −5%. Equation 3.11 is then employed to calculate the lower partial moments given at the bottom of the table. Effectively, only three observations (those shaded in the table) are used to calculate the downside risk measures as these are the only observations for which returns fall below −5%.

In contrast, a variance-based approach would use the full sample information. Although the problem of estimation error will not be addressed in detail until the next chapter (some comments with respect to the lpm method described so far can be found in Section 3.2.3.1), suspicion should already have been aroused by the superficiality of the method of estimating lower partial moments presented so far. For example, the same shortfall probability result would have been obtained for any threshold return between 0% and −13%.

3.2.2 Lower partial moments and multiple return series

So far, this section has estimated lower partial moments for a single return series, but how can we express the co-movements of assets in a lower partial moments framework? Will we be able to aggregate these risk measures in a portfolio with the same ease

Table 3.4 Example of lower partial moment calculation using Equation 3.11

R_i	d_i	$(R_i - \gamma)^0$	$(R_i - \gamma)^1$	$(R_i - \gamma)^2$
19.50	0	1	24.50	600.46
−33.43	**1**	**1**	**−28.43**	**808.07**
−33.43	1	1	−28.43	808.07
−13.29	**1**	**1**	**−8.29**	**68.80**
−13.29	1	1	−8.29	68.80
12.66	0	1	17.66	311.96
10.18	0	1	15.18	230.33
24.75	0	1	29.75	885.34
26.30	0	1	31.30	979.63
34.43	0	1	39.43	1554.60
7.73	0	1	12.73	162.01
12.61	0	1	17.61	310.02
−13.52	**1**	**1**	**−8.52**	**72.54**
−13.52	1	1	−8.52	72.54
4.89	0	1	9.89	97.77
21.62	0	1	26.62	708.56
11.79	0	1	16.79	281.92
0.38	0	1	5.38	28.99
43.34	0	1	48.34	2336.82
24.45	0	1	29.45	867.45
Calculated lpm risk measure for $m = 0$, 1 and 2				
		lpm$(\gamma, 0)$	lpm$(\gamma, 1)$	lpm$(\gamma, 2)$
		0.15	−2.26	47.47

as within classical portfolio theory? Starting from our definition of lower partial moments in Equation 3.11 and concentrating on downside variance,[30] we can extend the equation in a very natural way[31]

$$\mathrm{lpm}_{ij}(\gamma, 2) = \frac{1}{T}\sum_{i=1}^{T}(R_i - \gamma)(R_j - \gamma)d_i \tag{3.12}$$

Effectively, this means that Equation 3.12 will only use those occurrences where $d_i = 1$, ie, where asset i underperforms and does not achieve its threshold return. In the same way, we can define the co-lower partial moment of return on asset j, R_j, with asset i, R_i, thus

$$\mathrm{lpm}_{ij}(\gamma, 2) = \frac{1}{T}\sum_{i=1}^{T}(R_i - \gamma)(R_j - \gamma)d_j \tag{3.13}$$

Table 3.5 Co-lower partial moments calculated for two uncorrelated return series

Series 1		Series 2			
R_i	d_i	R_j	d_j	Eqn 3.12	Eqn 3.13
19.5	0	−27.8	1	0.0	−558.1
16.52	0	−27.9	1	0.0	−493.7
−33.43	1	−19.2	1	403.6	403.6
23.95	0	−35.4	1	0.0	−879.9
−13.29	1	−3.9	0	−9.2	0.0
12.66	0	21.0	0	0.0	0.0
10.18	0	30.1	0	0.0	0.0
24.75	0	28.9	0	0.0	0.0
26.30	0	43.8	0	0.0	0.0
34.43	0	−12.6	1	0.0	−297.7
7.73	0	11.2	0	0.0	0.0
12.61	0	33.7	0	0.0	0.0
3.10	0	−14.3	1	0.0	−75.1
−13.52	1	37.6	0	−363.4	0.0
4.89	0	35.9	0	0.0	0.0
21.62	0	11.0	0	0.0	0.0
11.79	0	−2.8	0	0.0	0.0
0.38	0	−8.1	1	0.0	−16.4
43.34	0	5.2	0	0.0	0.0
24.45	0	−23.9	1	0.0	−557.6

Equation 3.12: $(R_i - \gamma)(R_j - \gamma)d_i$. Equation 3.13: $(R_i - \gamma)(R_j - \gamma)d_j$.

In general (apart from lucky coincidences), we find that $d_i \neq d_j$, which means that we cannot come up with a symmetric co-lower partial moment matrix using the definition in Equations 3.12 and 3.13. An example of this can be seen in Table 3.5, where the numbers have been generated from two uncorrelated normal distributions with an average return of 10% and a volatility of 20% each.

We calculate the lower partial moments for assets *i* and *j* (and vice versa) according to Equations 3.12 and 3.13 and get

$$\begin{aligned}\mathrm{lpm}_{ij}(\gamma, 2) &= \frac{1}{T}\sum_{i=1}^{T}(R_i - \gamma)(R_j - \gamma)d_i \\ &= \frac{1}{20}(403.6 - 9.2 - 363.4) \\ &= 1.6\end{aligned}$$

$$\begin{aligned}\mathrm{lpm}_{ij}(\gamma,2) &= \frac{1}{T}\sum_{i=1}^{T}(R_i-\gamma)(R_j-\gamma)d_j\\ &= \frac{1}{20}(-558.1-\cdots\\ &\qquad -292.2-297.7-75.1-16.4-557.6)\\ &= -123.7\end{aligned}$$

The estimates are highly asymmetric (they differ substantially) and prone to the same estimation error problem that we have already seen with lower partial moments for a single asset. This asymmetry makes the application of standard (critical line) portfolio optimisers impossible.

However, with the advent of more general optimisers we can optimise (ie, minimise) lower partial moments in a simple Excel spreadsheet using the following short procedure:

Step 1. Fill *k* columns, where *k* is the number of assets, and *T* rows, where *T* is number of return observations or time periods, with the returns for the respective assets. Returns could either be historical data or obtained from fitted and bootstrapped distributions. Make the number of observations large so there are as few gaps in the distribution as possible. (See the recommended method in Section 3.2.3.2 for how to estimate lower partial moments correctly.)

Step 2. Calculate portfolio returns for each time period (row) for a given weight vector and calculate the lower partial moment of choice on the series of portfolio returns using the steps shown in Table 2.4.

Step 3. Use an optimiser to minimise lower partial moments, subject to a return constraint (and other constraints of choice).[32]

Step 4. Repeat Step 3 for different target returns to trace out an efficient frontier.

Fortunately, most properties we know from traditional mean–variance optimisation carry over to the lower partial moments framework. It has been shown that a continuous frontier exists, but also that the usual two-fund separation properties hold (the frontier becomes a straight line and we can separate preferences from relative asset weights) as soon as we include cash in the universe.[33]

However, this still leaves the problem that lower partial moments are not symmetric. To circumvent this, a symmetric lower partial moment measure has been described[34] that allows one to construct a symmetric lower partial moment matrix.[35] This procedure takes the individual lower partial moments and "glues" them together using ordinary correlations

$$\mathrm{lpm}_{ij}^{\text{symmetric}}(\gamma, 2) = \sqrt{\mathrm{lpm}_i(\gamma, 2)}\sqrt{\mathrm{lpm}_j(\gamma, 2)}\,\rho_{ij} \tag{3.14}$$

This allows us to calculate the lower partial moment of the portfolio as the usual quadratic form

$$\mathrm{lpm}_p(\gamma, 2) = \sum_{i=1}^{k}\sum_{j=1}^{k} w_i w_j \,\mathrm{lpm}_{ij}^{\text{symmetric}}(\gamma, 2) \tag{3.15}$$

Because this procedure reduces estimation error, it is more robust and will therefore improve out-of-sample performance.[36]

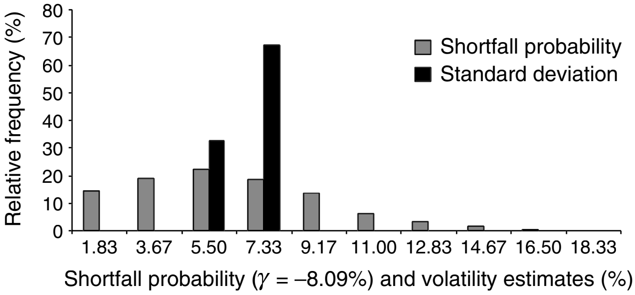
3.2.3 How to estimate lower partial moments correctly

Lower partial moments can be estimated the right or the wrong way. Unfortunately, a lot of criticism of lower partial moments as a concept has been caused by practitioners misguidedly using the incorrect estimation method, ie, utilising empirical return distributions.[37] However, we should not confuse practical matters of implementation (the problem of estimation) with conceptual issues. Hence, this section does not deal with conceptual concerns (for these, see Section 3.3) but instead focuses on implementation.[38]

3.2.3.1 The wrong way, or how to maximise estimation error

The flawed method of estimating lower partial moments, using the empirical distribution to calculate Equation 3.11, has been considered in some detail in Sections 3.2.1 and 3.2.2. As lower partial moments are based on the left tail of a distribution – the downside of asset returns – and as extreme realisations are by definition rarely observed, we suspect that estimation error might be a problem when lower partial moments are used. We also suspect that estimation error increases the more extreme the threshold we choose. A larger estimation error than given by variance-based risk measures would be a serious handicap to implementation as portfolio optimisation algorithms tend to leverage estimation error by taking on extreme positions. To appreciate the potential for estimation error in

Figure 3.6 Dispersion of estimates of shortfall measure, lpm(γ, m), and standard deviation for asset discussed in text; greater dispersion of shortfall measure implies greater potential for estimation error



lower partial moment estimates we will look at a simple sampling experiment.

Suppose we have an asset with a monthly mean of 1% and a monthly volatility of 5.77%. We want to simulate the dispersion of our estimates of shortfall probability, lpm$(\gamma, 0)$, and standard deviation (ie, volatility). To reflect the dispersion of both correctly in a standard histogram (showing estimates of the same magnitude), we choose the threshold value, γ, so that the true probability of shortfall equals the true volatility. Choosing a threshold value of −8.09%, we get (see Appendix A (on page 102) for numerical interpretation of expressions like 3.16)

$$\text{lpm}(-8.09\%, 0) = \int_{-\infty}^{-8.09\%} (R - \gamma)^0 f(R)\, \mathrm{d}R = 5.77\% \qquad (3.16)$$

We now draw 60 monthly observations, estimate shortfall probability and standard deviation and repeat this procedure 1,000 times. The histogram of estimates is plotted in Figure 3.6 and shows the much wider dispersion of the shortfall probability estimates. This confirms the belief that estimation error is a potentially serious problem for the implementation of lower partial moments methods.

3.2.3.2 Best practice

We have seen that using the (discrete) empirical distribution gives flawed results. Using instead the volatility estimates and calculating the shortfall probability using Equation 3.16 would obviously have led to much less diverse results. We recommend a similar route for estimating lower partial moments. Essentially, this is a multi-step procedure that can be automated:[39]

Step 1. Specify the candidate distributions to be fitted, making sure that they are consistent with the possible range of return realisations.

Step 2. Estimate all candidate distributions and find the distribution that best fits the data empirically, ie, determine the distribution function $\hat{f}(R)$ using appropriate methods.

Step 3. Integrate the best-fit distribution found in Step 2 to calculate the desired lower partial moment (see Appendix A (on page 102) for suitable integration techniques)

$$\operatorname{lpm}(\gamma, m) = \int_{-\infty}^{\gamma} (R - \gamma)^m \hat{f}(R)\, \mathrm{d}R \tag{3.17}$$

Two points with regard to Step 1. First, clearly it would be imprudent to consider fitting a normal distribution to a guaranteed fund product (like a protective put) as negative three-sigma events will not be feasible – put protection will in most cases not allow negative returns of this size. Second, although maximum likelihood methods are widely used, we will focus on the merits of fit-based estimation methods as they are significantly easier to grasp.[40] Moreover, they can be implemented in a straightforward way using accessible spreadsheet technology. These methods maximise the similarity between the empirical cumulative probability of the data and the theoretical cumulative probability of the fitted distribution.

One merit of fit-based methods is that they use the distribution function of the data directly. This is done in the next three steps:

Step 4. Rank all observations in a return series in ascending order, starting with the smallest realisation. Extract the empirical cumulative distribution by calculating $F(R_i) = i/T$, where i is the rank of the ith observation and T is the total number of observations.

Step 5. Given a set of starting values for the distributional parameters in our candidate distribution, calculate the distance between the empirical and the theoretical (fitted) cumulative distribution function. The most common procedure is to calculate the absolute distance $|F(R_i) - \hat{F}(R_i)|$, but other penalty functions are also applicable. Calculate the total variation.

Step 6. Repeat Step 5 until the total variation is minimised or, alternatively, until the Kolmogorov–Smirnov statistic, D, given by[41]

$$D = \max_i (D_i) = \max_i (|F(R_i) - \hat{F}(R_i)|) \tag{3.18}$$

is minimised to find the best-fitting combination of parameters.

Note, however, that Equation 3.18 is determined by the largest distance, so it might rank a distribution that does not fit in general above a distribution that fits very well apart from a single data point – which essentially is the logic of any min–max criterion (choosing the minimum difference over a range of maximums). Additionally, since the vertical difference between $F(R_i)$ and $\hat{F}(R_i)$ itself has a distribution

$$\hat{\sigma}^2(D_i) = \frac{\hat{F}(R_i)[1 - \hat{F}(R_i)]}{T}$$

and since it becomes more volatile as we come closer to the centre of the distribution, the Kolmogorov–Smirnov statistic will pay considerably less attention to fitting the tails, which is potentially hazardous for lower partial moment estimation.

Step 6*. An alternative to the Kolmogorov–Smirnov statistic is the Anderson–Darling statistic, A, calculated as[42]

$$A^2 = \int_{-\infty}^{+\infty} \frac{[F(R_i) - \hat{F}(R_i)]^2}{\hat{\sigma}^2} \hat{f}(R)\, \mathrm{d}R \tag{3.19}$$

This puts less emphasis on the highly variable values (scaling them down by their respective variance) and more emphasis on the more likely values (weighting distances by their relative importance).

The results of an example calculation for S&P 500 Index data are shown in Figure 3.7. The fitted logistic distribution, which has an Anderson–Darling statistic value of 0.14, can be written as

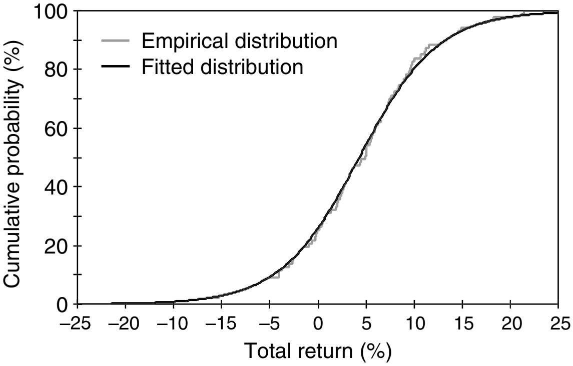
$$\hat{f}(R) = \frac{z}{\beta(1+z)^2}, \qquad \hat{F}(R) = \frac{1}{1+z}$$

where

$$z = \exp\left(-\frac{R-\alpha}{\beta}\right), \quad \alpha = 0.043028, \quad \beta = 0.040792$$

Step 7. The last step is numerical integration of the fitted distribution function. As we already know $\hat{F}(R_i)$, the evaluation of

Figure 3.7 Fit of logistic distribution to S&P 500 Index data using the Anderson–Darling statistic (Step 6* in text)



Data is S&P 500 Index quarterly total returns from second quarter of 1979 to fourth quarter of 2000. *Source*: DataStream.

a target semivariance (target rate of 0%) becomes much easier; numerical integration essentially means multiplying the probability of falling into a bucket by the average value of that bucket (see Appendix A on page 102). Hence, the lower partial moment integral can be approximated by

$$\int_{-\infty}^{0} \hat{f}(R)R^2 \,\mathrm{d}R = \sum_i [\hat{F}(R_{i+1}) - \hat{F}(R_i)]\left[\frac{R_{i+1} + R_i}{2}\right] = 3.17\%$$

As the difficulty of the approach recommended here is to find the best-fitting distribution (a set of candidate distributions has to be estimated and the best-fitting distribution has to be selected using statistical criteria) and because these distributions change (causing a stability problem), it has also been proposed that a more continuous – ie, less arbitrary – distribution should be created by repeatedly sampling from the empirical distribution (known as bootstrapping). Typically, we would use monthly data and sample 12 realisations to create a single annual return. Repeating this procedure 1,000 times would give a smooth distribution of annual returns. However, doing this would also destroy the non-normality in the data as the independent draws generated by the bootstrapping procedure would generate a (log)normal distribution – an instance of the Central Limit Theorem at work again. Hence, this procedure is not recommended.

3.3 A COMPARISON OF LOWER-PARTIAL-MOMENTS-BASED AND VARIANCE-BASED METHODS IN PORTFOLIO CHOICE

So far, this chapter has critically reviewed the assumption of normality of asset returns, introduced downside risk measures and described the estimation problems that arise when these measures are used. In this section we want to comment on the use of downside risk measures in portfolio optimisation compared to Markowitz optimisation.

There is fierce debate between supporters and critics of classical or mean–variance based portfolio theory.[43] One battleground is utility theory: is a preference for downside risk compatible with rational decision-making (ie, expected utility maximisation), and are such preferences plausible?[44] How well do mean–variance based solutions approximate expected utility?[45] Another area of contention is actual portfolio allocations: how different are portfolios constructed using the two approaches? From a practical point of view, the last question is by far the most interesting as it is very difficult to contradict a client who claims to have nonmean–variance preferences. Hence, we will concentrate on how variance based portfolios compare with portfolios constructed using lower partial moments.

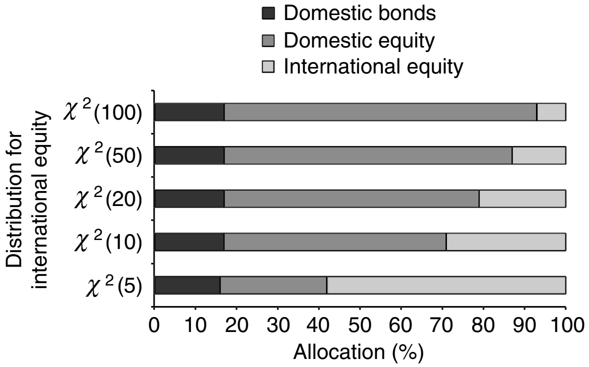
Introducing lower partial moments into portfolio optimisation merely requires a change of risk measure. Instead of minimising variance, we minimise $\operatorname{lpm}(\gamma, m)$ with respect to the vector of portfolio weights, subject to a return constraint, a budget constraint and whichever other constraints require

$$\left.\begin{aligned} \min_{w} \operatorname{lpm}(\gamma, m) \\ \boldsymbol{w}'\boldsymbol{R} = \bar{R} \\ \boldsymbol{w}'\boldsymbol{I} = 1 \\ \vdots \end{aligned}\right\} \qquad (3.20)$$

Changing the required return, we can trace out an efficient frontier, ie, plot the portfolios that result from Equation 3.20.

Suppose that we want to investigate how skewness in asset returns affects the optimal portfolio allocations obtained from Equation 3.20 and how these allocations compare with mean–variance based portfolios. A three-asset example will demonstrate the issues. We will assume, for illustration only, that domestic bonds (5% volatility) and domestic equity (15% volatility) are normally distributed,

Figure 3.8 Effect of skewness in asset returns on portfolio allocations obtained using lower partial moments (Equation 3.20)



Plot shows effect of diminishing positive skewness in international equity distribution ($\chi^2(\nu)$) on overall asset allocation. $\chi^2(100)$ distribution approximates normal, ie, bar represents a mean–variance based allocation. See text for further explanation.

whereas international equity (20% volatility) has a skewed distribution with various degrees of skewness. Specifically, international equity is assumed to be distributed as $\chi^2(\nu)$, where ν denotes the different degrees of freedom.

Samples of correlated returns (10,000 observations each) are created using the fact that $\chi^2(\nu)$ has mean ν and standard deviation $\sqrt{2\nu}$.[46] This allows us to model international equity returns with the desired mean and volatility. Although $\chi^2(\nu)$ is positively skewed (a desirable trait for lower partial moments), it will approach the normal distribution as $\nu \to \infty$.[47] This means that we can observe how a continuous shift from skewed distributions to a normal distribution affects portfolio composition. Correlations between variables are maintained by using the rank order correlation method.[48] For each feasible weight vector we calculate the risk characteristics using the series of 10,000 portfolio return observations.

Figure 3.8 shows the result of this experiment. We set a 7% annual return while minimising downside volatility with respect to a zero-loss return target, ie, $\text{lpm}(0,2)$. For very positively skewed international equity returns (small ν), this asset class dominates the allocation. This is different from mean–variance based solutions – approximated in Figure 3.8 by the bar for $\chi^2(100)$ – which fail to detect the desired feature of positive skewness and so prefer the asset with lower volatility (domestic equity), which is actually riskier. As international equity returns approach normality both solutions coincide. This is not surprising as, in a normally distributed world, a portfolio with low downside risk is, in fact, merely a portfolio with low volatility. If, for example, investors were concerned with the probability of loss rather than the volatility of returns, the set of efficient solutions would not change; as long as returns are normally distributed we will find that, for any given return, the portfolio with the lowest probability of loss is also the one with minimum volatility. Hence, the set of efficient portfolios stays the same (all are mean–variance efficient portfolios), although, depending on their loss-aversion, investors might choose different portfolios.

Conversely, if returns are significantly non-normal, solutions provided under the normality assumption will depart dangerously from the level of downside protection that could have been achieved if non-normality had been properly addressed. Non-normal returns will result in lower partial moment-minimised portfolios that dominate (plot above) mean–variance portfolios in mean lower partial moment space while being dominated in mean–variance space as a different objective is optimised.[49] The degree of non-normality and the investor's conviction about its stability will determine the choice of methodology.

To end this section, some brief comments should be made about how the choice of target return affects the difference between variance based and shortfall variance based (lpm-based) allocations. As the maximum return portfolio is composed from the maximum return assets, risk measures play almost no role in its construction. This changes as we move to the minimum risk portfolio, where assets are chosen purely on the basis of their riskiness. Hence, the difference between lpm-based and variance based portfolio construction will, other things being equal, increase as we move towards lower return requirements. A similar logic applies to the threshold return: the more it deviates to the left from the mean return, the more the allocations differ. We can summarise by saying that differences between the two methods increase with skewness, decrease with return requirements and decrease with threshold return.[50] To the extent that skewness can be predicted (see Section 3.1), lower partial moment based allocations are a powerful alternative to traditional portfolio construction.

3.4 SUMMARY

Lower partial moments offer a convenient way to model non-normality in returns data but still use an optimisation framework that is very much related to traditional Markowitz optimisation (and identical if symmetric co-moments are assumed). However, estimation error is a serious problem in implementation – much more so than with variance based measures. The problem can be partially overcome using the procedures reviewed in this chapter and suggested in the associated literature. Whether the modelling effort leads to solutions that are statistically different in-sample from traditional solutions depends critically on the degree of skewness. Whether these solutions add out-of-sample value depends on the persistence of deviations from non-normality. Dogmatic views of any persuasion are of little help.

Currently, most risk management software providers use cross-sectional models. We believe this is not necessarily the best choice for the following reasons:

1. Time series models are designed to explain the variation in asset returns, in other words, risk.[51] In contrast, cross-sectional models explain the variation in cross-sectional returns, in other words, conditional asset means. This is different to modelling risk and part of the legacy of cross-sectional models, which were first developed for the purpose of forecasting returns. The author believes that cross-sectional models are transformed alpha models and not original risk models. Fama and French (1993) themselves argue that a time series model is more appropriate for estimating risk.[52]
2. Time series models will clearly estimate individual beta exposure with (estimation) error: $\beta_{ij}^{\text{OLS}} = \beta_{ij}^{\text{true}} + \varepsilon_{ij}$. However, this does little harm at the portfolio beta level as it simply diversifies away when we add up individual beta exposures

   $$\beta_{ij}^{\text{portfolio}} = \frac{1}{n}\sum_{i=1}^{n}\beta_{ij}^{\text{true}} + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_{ij} = \frac{1}{n}\sum_{i=1}^{n}\beta_{ij}^{\text{true}}$$

   given that betas are individually unbiased (we assumed equal weighting of individual assets). This is in stark contrast to cross-sectional models. If individual betas (independent variables) are measured with errors, the resulting regression (cross section of returns against cross section of betas) coefficients, ie, the factor returns, are biased. However, errors in factor returns do not diversify. They affect the whole portfolio and are in the databank (of factor returns) forever.

3. Cross-sectional models are plagued with econometric problems. Regressions suffer from heteroskedasticity, ie, not all observations carry the same information. Observations with large errors should be weighted less but there is little available theory that could guide model builders. Also multivariate outliers are difficult to spot and might have a non-trivial effect on factor returns. Given that misestimated factor returns do not diversify away, some risk model providers use robust regressions instead. Robust regressions tend to move factor risks to residual risks. Given that residual risks diversify quickly this might lead to problems if the suspected outlier is not an outlier but an informative data point instead.
4. Cross-sectional models use unrealistic and extreme sector sensitivities that can only take values of zero or one. However, multiproduct firms have different exposures to different industries. If this is not correctly specified, the resulting estimates of factor returns will be misestimated and hence risk forecasts are likely to be wrong. Let us illustrate this point. The return of a given industry at a given point in time equals the regression coefficient on a dummy variable with unit exposure if the stock belongs to a given industry and zero else. These prespecified unitary industry exposures are most certainly wrong as stocks within an industry have different exposures to the risk factor running through an industry. Up and downstream firms in the oil industry, for example, will have different exposures (betas) to the oil factor, ie, they offer a different leverage on the oil industry. Providers of cross-sectional risk models try to mitigate this problem with finer and finer industry definitions and multi-industry assignments. However, the problem of measuring industry exposures using a scale that is predefined and discrete rather than dynamic and continuous cannot be fully addressed in cross-sectional models.
5. In cross-sectional models all stocks within a sector are generally assumed to respond identically to an industry shock.

In our view this makes cross-sectional models difficult to apply for hedging purposes. Every cash neutral allocation within an industry exhibits zero industry risk exposure in the logic of cross-sectional models, while a long–short portfolio of high/low industry beta portfolios will in reality exhibit sector risk.

6. Risk estimates from cross-sectional models depend on the industry universe used. This is why there is no one-size-fits-all cross-sectional model and why risk model providers seem to bias their estimation of actor returns to where their book of business is. In contrast, factor returns in a time series model are given, while beta exposures will be calculated on a stock-by-stock basis without any consideration to the wider universe.

However, these relative advantages come at a cost. Time series beta estimates might lack the responsiveness of cross-sectional models. Cross-sectional models have little problem with firms that change their capital structure overnight. In contrast, it might take a time series model some time to realise a change in sensitivities (betas). If true betas change quickly, a time series regression approach might be at a disadvantage as it averages over a (too) long history while prespecified betas are more responsive if correctly specified. In practice, these disadvantages are not meaningful for fast updating short- to medium-term risk models (two years of daily data or less) as opposed to long-term models (monthly updating over five years of data).

APPENDIX A: INTEGRATION TECHNIQUES FOR CALCULATING LOWER PARTIAL MOMENTS

There are three ways to calculate lower partial moments by integration after best-fitting a distribution to the historical data. The first and easiest way is to use a software package, like MATLAB, MAPLE or MATHEMATICA, that is capable of undertaking symbolic math calculations.

However, often we do not have a package like this at hand, we cannot integrate it into our other routines, or there is no simply no closed-form solution. In this case we have to numerically integrate the expression for the lower partial moment.

A1 Numerical integration

Suppose we need to calculate the shortfall probability for a normal distribution representing an asset with 3% average return and 6% volatility. For arbitrary continuous distributions we find that

$$\text{lpm}(0,0) = \int_{-\infty}^{0} f(R)\,\mathrm{d}R \tag{3.21}$$

In the present case, assuming normality, we already know that

$$f(R) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{R-\mu}{\sigma}\right)^2\right\}$$

which allows us to set up a spreadsheet as in Table 3.6.

We will use the trapezium rule to illustrate the concept of numerical integration.[53] The whole area under the bell-shaped curve is divided into vertical strips of equal width. This is done in the first column of Table 3.6, where we have chosen a width of 2% between two realisations of R_i. The smaller the width, the more precise our approximations will become. Note that our grid points (R_i) cover about three standard deviations from the mean in both directions $(-15\% = 3\% - 3 \times 6\% \leqslant R_R \leqslant 3\% + 3 \times 6\% = 21\%)$.

For each grid point we calculate the density $f(R_i)$, as in the second column of Table 3.6. Now the area under the bell curve between any two grid points can be approximated by a rectangle; for example, the area between −6% and −8% is approximated by a rectangle with a width of 2% and an average height of $(2.2 + 1.2)/2 = 1.7$. This value goes in the third column in the row for $R_i = -6\%$.

The fourth column contains the probabilities of a return falling between any two values. Continuing our example, the probability of a return falling between −6% and −8% is approximately

$$\underbrace{\tfrac{1}{2}|f(-6\%) - f(-8\%)|}_{\text{height}}\ \underbrace{2\%}_{\text{width}} = 3\%$$

This calculation is repeated for all buckets in the third column. Summing up the relevant entries, we get the shortfall probability according to

$$\begin{aligned} &\text{lpm}(0,0) \\ &\quad= \int_{-\infty}^{0} f(R)\,\mathrm{d}R \\ &\quad\approx \tfrac{1}{2}|f(-14\%) - f(-16\%)|2\% + \cdots + \tfrac{1}{2}|f(0) - f(-2\%)|2\% \\ &\quad= 30\% \end{aligned}$$

Table 3.6 Example of numerical integration

R_i(%)	$f(R_i)$	A	B	C
−16	0.0			
−14	0.1	0.08	0.00	−0.025
−12	0.3	0.21	0.00	−0.054
−10	0.6	0.46	0.01	−0.102
−8	1.2	0.94	0.02	−0.169
−6	2.2	1.70	0.03	−0.238
−4	3.4	2.76	0.06	−0.276
−2	4.7	4.03	0.08	−0.242
0	5.9	5.28	0.11	−0.106
2	6.6	6.21	0.12	0.124
4	6.6	6.56	0.13	0.393
6	5.9	6.21	0.12	0.621
8	4.7	5.28	0.11	0.740
10	3.4	4.03	0.08	0.726
12	2.2	2.76	0.06	0.608
14	1.2	1.70	0.03	0.442
16	0.6	0.94	0.02	0.281
18	0.3	0.46	0.01	0.158
20	0.1	0.21	0.00	0.078

A, $\frac{1}{2}|f(R_i) - f(R_{i-1})|$.

B, $\frac{1}{2}|f(R_i) - f(R_{i-1})|\Delta R$.

C, $\frac{1}{2}|f(R_i) - f(R_{i-1})|\Delta R \cdot \frac{1}{2}(R_i + R_{i-1})$.

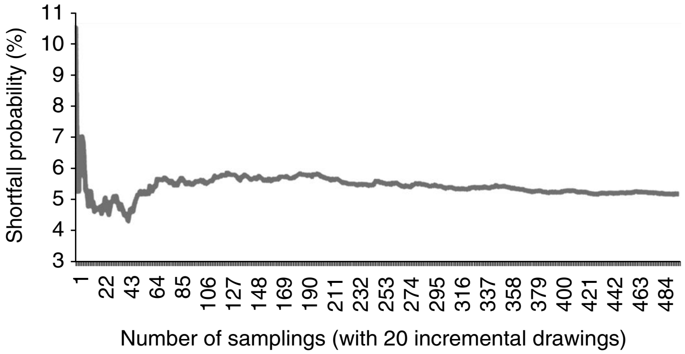
Obviously, the entries in the fourth column have to sum up to 1, ie, the area under the curve is 1.

We can extend our analysis by calculating the average conditional loss[54]

$$\text{average conditional loss} = \frac{\int_{-\infty}^{0} f(R)R\,\mathrm{d}R}{\int_{-\infty}^{0} f(R)\,\mathrm{d}R} \tag{3.22}$$

Practitioners sometimes do this when moving from domestic to global stocks to give clients a feel for risk. The average conditional loss adds up all probability-weighted losses. However, as we want to calculate the average conditional loss (ie, the average loss after a loss took place), we have to standardise (divide) by the probability of making a loss (already known to be 30%). The fifth column in Table 3.6 numerically integrates $\int_{-\infty}^{0} f(R)R\,\mathrm{d}R = -1.21\%$. Hence,

Figure 3.9 Convergence for Monte Carlo integration



the average conditional loss becomes

$$\text{average conditional loss} = \frac{-1.21\%}{30\%} = -4.04\%$$

A2 Monte Carlo integration

The second method considered here is Monte Carlo integration. Essentially this is a "brute force" method – not elegant, but it does work if computing time is inexpensive.

The process is as follows. Take a limited number of random draws from the estimated distribution. Calculate the desired lower partial moment, or any other statistic – like that in Equation 3.22. Increase the data set by sampling further from the estimated distribution and plot the calculated risk figure. As the window of sampled data increases the estimate will start to stabilise. How fast the empirical values converge to the true fixed value depends on which statistic is calculated. The further out into the tail we are looking, the longer convergence takes.

The result of this process (mirroring the assumptions in Section 3.2.3) can be seen in Figure 3.9. Depending on what criteria are set for convergence, it will take about 250 draws of 20 observations each, ie, 5,000 data points, until the estimate converges (stabilises). The first 20 data points result in a shortfall probability of 10.5%. After another 50 drawings, each with 20 observations, this value drops to about 5%, but many more drawings are required for further convergence.

EXERCISES

Suppose we are given $i = 1, \dots, n$ of i.i.d. returns r_i from which we can calculate a performance measure. Let us denote by θ the true but unknown performance measure, $\hat{\theta}$ its empirical estimate and θ^* its resampled value. The returns data is assumed to be representative of the true but unknown cumulative return distribution. Taking $b = 1, \dots, B$ bootstraps of the performance measure θ_b^* (sample n draws from the empirical return distribution with replacement and calculate a new value for a chosen performance measure) provides us with an empirical estimate of the unknown sampling distribution that will be in many cases impossible to derive. This is the basic idea of bootstrapping. Shape and dispersion of θ^* around $\hat{\theta}$ will mimic shape and dispersion of $\hat{\theta}$ around θ. As we have made no further assumptions, bootstrapping does free us from simple models and restrictive distributional assumptions.

1. Download a series of hedge fund returns from the internet (for example, the website of EDHEC) and calculate return/risk ratios, where risk is given from lower partial moment estimates for $m = 1, 2, 3, 4$ and $g = 0$ and return equals sample means.

2. Bootstrap the different lower partial moments. Which has the highest estimation error?

3. Use bootstrapping to determine the statistical significance of your performance measures. Explain your results.

[1] See Mandelbrot (1963) or Fama (1965) for early references and Duffie and Pan (1997) for a more recent review.

[2] See Jorion (2001, p. 104).

[3] See Bamberg and Dorfleitner (2001) for a review.

[4] See Bekaert *et al* (1998) for a review of distributional characteristics for emerging market debt.

[5] However, it should be noted that high-frequency returns in developed markets (hourly or daily returns) also exhibit considerable non-normality (much higher than monthly returns), so this may be fallacious for those with short time horizons.

[6] See Bera and Jarque (1987).

[7] An alternative test for the significance of higher moments is to take the deviations from the mean return, raise them to the power of three or four (also accounting for normal kurtosis) and regress them against a constant. The t value on the constant measures significance.

[8] While studies by Singleton and Wingender (1986) or Peiro (1994, 1999) show no persistence in skewness, others, such as Nawrocki (1991), have been more optimistic.

[9] This test is taken from Kahn and Stefek (1996).

[10] For a symmetric distribution this term should be zero.

[11] This is the so-called Lindberg–Levy form. The Lindberg–Feller version is even stronger as it states that the sum of random variables is normally distributed even if they all come from different distributions. See Green (2000, p. 116).

[12] If the sum of one-period logarithmic returns, $\sum \ln(1 + R_i)$, is normal (CLT), the multi-period return $\exp\{\sum \ln(1 + R_i)\}$ is lognormal (a random variable is lognormal if its log is normal).

[13] See Bookstaber and Langsam (1988) as a classic reference.

[14] See Siegel (2000).

[15] The proposed simulation procedure ignores any intertemporal dependence that might exist in the data.

[16] As convergence by definition takes time, most investors do not have the luxury to wait until convergence occurs; even long-term investors have short-term reporting horizons that make the application of the CLT less applicable.

[17] Garch is the abbreviated form of "generalised autoregressive conditional heteroscedasticity". An up-to-date review of Garch modelling can be found in Alexander (2001, Chapter 4).

[18] See Mills (1999, p. 186).

[19] See the S-Plus code from McNeil (2001), which produces hill plots as well as more sophisticated estimators for tail indexes.

[20] See Hall (1982).

[21] An unexpected feature of fat-tailed distributions is that large multi-period losses grow with $T^{1/\alpha}$ rather than with $T^{1/2}$ for the normal distribution. For large enough time horizons the fat-tailed distributions ($\alpha > 2$) will show lower probabilities of large multi-period losses. See Embrechts (2000) and, in particular, Danielson and DeVries (2000, p. 91).

[22] See Kim and Finger (2000) for a very recent application to investigate the breakdown of correlation between markets.

[23] This section draws on Hamilton (1994, p. 684).

[24] We could also test the significance of the mixture of normals relative to the unconditional normal distribution, as suggested in Kim and Finger (2000), using a likelihood ratio test. The test statistic is given as

$$-2c \ln\left(\frac{L_{\text{constrained}}}{L_{\text{unconstrained}}}\right) \sim \chi^2(4), \quad c = \frac{N-3}{N}$$

where N is the number of observations and $L_{\text{constrained}}$ denotes the value of the constrained likelihood function (where the parameters are equal between regimes). The value of the test statistic is 12.16, against a critical value of 9.48 (for the 5% confidence level). Hence, the mixture-of-distributions model is statistically different from its simple alternative.

[25] See Stutzer (2000).

[26] As *ex ante* and *ex post* should not be different in large samples, we use an equals sign in Equation 3.11, even though this is technically incorrect, as E is the *ex ante* expectations operator.

[27] Note that this is different from the conditional underperformance, which is the average underperformance if underperformance occurs.

[28] If we set $\gamma = \bar{\mu}$ (where the threshold equals the distribution mean), we arrive at a measure called "semivariance" because it calculates the variance of return realisations below the mean return

$$\operatorname{lpm}(\bar{\mu}, 2) = \frac{1}{T}\sum_{i=1}^{T}(R_i - \bar{\mu})^2 d_i, \quad d_i = \begin{cases} 0, & R_i > \bar{\mu} \\ 1, & R_i \leqslant \bar{\mu} \end{cases}$$

However, for every $d_i = 1$, we know that $1 - d_i = 0$. It is now easy to show that

$$\begin{aligned} \operatorname{lpm}(\bar{\mu}, 2) + \operatorname{upm}(\bar{\mu}, 2) &= \frac{1}{T}\sum_{i=1}^{T}(R_i - \bar{\mu})^2 d_i + \frac{1}{T}\sum_{i=1}^{T}(R_i - \bar{\mu})^2(1 - d_i) \\ &= \frac{1}{T}\sum_{i=1}^{T}(R_i - \bar{\mu})^2 = \sigma^2 \end{aligned}$$

In the case of symmetric distributions this term simplifies even further and we get

$$2\,\mathrm{lpm}(\bar{\mu}, 2) = \sigma^2$$

which can be used as a simple normality check. If the variance is roughly twice the semivariance, it is close to symmetry (symmetry does not imply normality, though, as distributions might be symmetric but still exhibit fat tails (see Section 3.1)).

29 Higher-order lower partial moments can be interpreted as shortfall skewness, kurtosis, etc.

30 A more general formulation for lower partial co-movements would write lower partial comoments as

$$\mathrm{lpm}_{ij}(\gamma, m) = \frac{1}{T}\sum_{i=1}^{T}(R_i - \gamma)^{m-1}(R_j - \gamma)d_i$$

31 See Harlow (1991). It is worth mentioning that Markowitz (1959) was the first to introduce downside risk measures into the portfolio management arena. However, as he used a slightly different concept than the current mainstream definition in Equation 3.11, it is not followed in this context.

32 Examples of such optimisers include Palisade Evolver and Frontline Solver. Note that this does not necessarily lead to the optimal solution (as described in Chapter 6) and is effectively close to scenario optimisation.

33 See Hogan (1974) for a review of semivariance and Bawa (1978) for a generalisation of all lower partial moments.

34 Nawrocki (1991) deserves credit for introducing symmetric lower partial co-movements.

35 The cause of asymmetry in Table 3.4 was estimation error. It is true that symmetry is a technically convenient property, but it is not one that we would expect economically.

36 See Nawrocki (1991, p. 470), who modelled the out-of-sample performance of alternative risk measures.

37 See Sortino and Forsey (1996) versus Kaplan and Siegel (1994).

38 Rom and Ferguson (1993, 1994) can be seen as a typical example of a heated conceptual debate. A more balanced view can be found in Kahn and Stefek (1996).

39 Rather than fitting the whole distribution, it is possible to use the Fisher and Cornish approximation (see Hull 2000, p. 366) to calculate the percentiles of a distribution merely by calculating skewness and kurtosis. We know from basic statistics that the α-percentile of a normal distribution is given as $\mu - z_\alpha \sigma$, where z_α represents the α-percentile of a standard normal distribution. The Fisher and Cornish approximation will "adjust" z_α to reflect non-normality. In the case of skewness only we get $\mu - z_\alpha^* \sigma$, with $z_\alpha^* = z_\alpha + \frac{1}{6}(z_\alpha^2 - 1)$ skew. If the skew is negative (longer left tail), z_α^* will be even more negative (the result of adding a negative number to an already negative number).

40 See Vose (2000) for an excellent review of these techniques.

41 Further references can be found in Vose (2000), who also gives a concise description of distribution fitting.

42 See Anderson and Darling (1952).

43 See the debate between Rom and Ferguson (1993, 1994) and Kaplan and Siegel (1994) in the *Journal of Investing*, where business interests meet dogmatics.

44 See Fishburn (1997) and Bawa (1978), who show that dominance of mean- and lower partial moment based decision rules is a necessary condition for expected utility maximisation (despite not showing sufficiency).

45 See Levy and Markowitz (1979) or Fung and Hsieh (1997).

46 We also assume that correlation between domestic bonds and domestic equity is 0.2 and that between domestic bonds and international equity it is 0.1, while equities are correlated with 0.7. Bond returns are expected to be 4.5%, while equity markets are supposed to deliver 7.5% annual returns. The caveat here is the assumption of the presence of positive skewness, which comes without penalty in expected returns as investors would bid up the prices of these assets, lowering the expected returns. Issues like this are dealt with in the literature on asset pricing using higher moments. See Arditti and Levy (1976) and Krauss and Litzenberger (1976) or Lai (1991) and Chunhachinda *et al* (1997) for more recent studies.

47 However, the distribution will not become perfectly symmetric as the normal distribution peaks at its mean while $\chi^2(\nu)$ peaks at $\nu - 2$.

48 Simplified, this means to correlate ranks of the individual distributions rather than continuous realisations. See Iman and Conover (1982).

49 Mean lower partial moment space refers to a two-dimensional graph where the x-axis charts the calculated lower partial moments and the y-axis represents portfolio mean.

50 This also means that these differences will tend to decrease with the number of assets, as portfolio skewness should be diversified away, and that they are probably less relevant to active risks (tracking error) as the possibility of long and short positions should diversify skewness away even faster.

51 See Rosenberg (1974) as the first reference for multifactor risk models.

52 Another way of saying this is that factor returns from cross-sectional models are actually market implied factor returns and better represent what the market might have thought about a factor return rather than what it truly was. Suppose we calculate the betas of all stocks with respect to oil price changes and use these betas as a stock characteristic. Regressing this characteristic against next period stock returns yields a market implied return on oil that will generally differ from the realised return on oil.

53 See Ostajewski (1990) for an introduction to numerical integration. A spreadsheet-based treatment of this can be found in Mesina (2001).

54 See Jorion (2001, p. 97).

REFERENCES

Alexander, C., 2001, *Market Models* (Chichester: John Wiley & Sons).

Anderson, T., and D. Darling, 1952, "Asymptotic Theory of Certain Goodness of Fit Criteria Based on Stochastic Processes", *Annals of Mathematical Statistics* 23, pp. 193–212.

Arditti, F., and H. Levy, 1976, "Portfolio Efficiency Analysis in Three Moments: The Multiperiod Case", *Journal of Finance* 30, pp. 797–809.

Bamberg, G., and G. Dorfleitner, 2001, "Fat Tails and Traditional Capital Market Theory", Working Paper, University of Augsburg.

Bawa, V., 1978, "Safety First, Stochastic Dominance and Optimal Portfolio Choice", *Journal of Financial and Quantitative Analysis* 13, pp. 255–71.

Bekaert, G., C. Erb, C. Harvey and T. Viskanta, 1998, "Distributional Characteristics of Emerging Markets Returns and Asset Allocation", *Journal of Portfolio Management* 24, pp. 102–15.

Bera, A., and C. Jarque, 1987, "A Test for Normality of Observations and Regression Residuals", *International Statistical Review* 55, pp. 163–72.

Bookstaber, R., and J. A. Langsam, 1988, "Portfolio Insurance Trading Rules", *Journal of Futures Markets* 8(1), pp. 15–32.

Chunhachinda, P., S. Dandapani, S. Harnid and A. Prakash, 1997, "Portfolio Selection and Skewness: Evidence from International Stock Markets", *Journal of Banking and Finance* 21, pp. 143–67.

Danielson, J., and C. DeVries, 2000, "Value at Risk and Extreme Returns", in P. Embrechts (ed), *Extremes and Integrated Risk Management* (London: Risk Books), pp. 85–106.

Duffie, D., and J. Pan, 1997, "An Overview over Value at Risk", *Journal of Derivatives* 4(3), pp. 7–49.

Embrechts, P., 2000, *Extremes and Integrated Risk Management* (London: Risk Books).

Fama, E. F., 1965, "The Behaviour of Stock Market Prices", *Journal of Business* 38, pp. 34–105.

Fama, E. F., and K. R. French, 1993, "Common Risk Factors in the Returns of Stocks and Bonds", *Journal of Financial Economics* 33(1), pp. 3–56.

Fishburn, P., 1997, "Mean Risk Analysis with Risk Associated with Below-Target Returns", *American Economic Review* 67(2), pp. 116–26.

Fung, W., and D. Hsieh, 1997, "Is Mean Variance Analysis Applicable to Hedge Funds?", Working Paper, Duke University.

Green, W., 2000, *Econometric Analysis*, Fourth Edition (New York: Prentice-Hall).

Hall, P., 1982, "On Some Simple Estimates of an Exponent of Regular Variation", *Journal of the Royal Statistical Society* 44, pp. 37–42.

Hamilton, J., 1994, *Time Series Analysis* (Princeton, NJ: Princeton University Press).

Harlow, W., 1991, "Asset Allocation in a Downside Risk Framework", *Financial Analysts Journal* 47, pp. 28–40.

Hogan, W., 1974, "Toward the Development of an Equilibrium Based Capital Market Model Based on Semivariance", *Journal of Financial and Quantitative Analysis* 9, pp. 1–11.

Hull, C., 2000, *Options, Futures and Other Derivatives*, Fourth Edition (Englewood Cliffs, NJ: Prentice Hall).

Iman, R., and W. Conover, 1982, "A Distribution Free Approach to Inducing Rank Order Correlation among Input Variables", *Communications in Statistics* 11, pp. 311–34.

Jorion, P., 2001, *Value at Risk*, Second Edition (New York: McGraw-Hill).

Kahn, R., and D. Stefek, 1996, "Heat, Light and Downside Risk", Barra Research Paper.

Kaplan, P., and L. Siegel, 1994, "Portfolio Theory Is Alive and Well", *Journal of Investing* 3, pp. 18–23.

Kim, J., and C. Finger, 2000, "A Stress Test to Incorporate Correlation Breakdown", Riskmetrics Paper.

Krauss, A., and R. Litzenberger, 1976, "Skewness Preference and the Valuation of Risky Assets", *Journal of Finance* 31, pp. 1085–1100.

Lai, T., 1991, "Portfolio Selection with Skewness, a Multiple Objective Approach", *Review of Quantitative Finance and Accounting* 1, pp. 293–305.

Levy, H., and H. Markowitz, 1979, "Approximating Expected Utility by a Function of Mean and Variance", *American Economic Review* 69, pp. 308–17.

McNeil, A., 2001, "EVIS (S-Plus Code for Extreme Value Theory)", URL: http://www.math.ethz.ch/~mcneil.

Mandelbrot, B., 1963, "The Variation of Certain Speculative Prices", *Journal of Business* 36, pp. 394–419.

Markowitz, H., 1959, *Portfolio Selection: Efficient Diversification of Investments* (New Haven: Yale University Press).

Mesina, M., 2001, *Numerische Mathematik mit Excel* (Poing: Franzis).

Mills, T., 1999, *The Econometric Modelling of Financial Time Series*, Second Edition (Cambridge University Press).

Nawrocki, D., 1991, "Optimal Algorithms and Lower Partial Moment: *Ex Post* Results", *Journal of Applied Economics* 23, pp. 465–70.

Ostajewski, A., 1990, *Advanced Mathematical Models* (Cambridge University Press).

Peiro, A., 1994, "The Distribution of Stock Returns: International Evidence", *Journal of Applied Financial Economics* 4, pp. 431–9.

Peiro, A., 1999, "Skewness in Financial Returns", *Journal of Banking and Finance* 23, pp. 847–62.

Rom, B., and K. Ferguson, 1993, "Post Modern Portfolio Theory Comes of Age", *Journal of Investing* 2, pp. 11–17.

Rom, B., and K. Ferguson, 1994, "Portfolio Theory Is Alive and Well: A Response", *Journal of Investing* 2, pp. 24–44.

Rosenberg, B., 1974, "Extra Components of Covariance in Security Returns", *Journal of Financial and Quantitative Analysis* 9, pp. 263–74.

Siegel, J., 2000, *Stocks for the Long Run* (New York: McGraw-Hill).

Singleton, J., and J. Wingender, 1986, "Skewness Persistence in Common Stock Returns", *Journal of Financial and Quantitative Analysis* 21, pp. 335–41.

Sortino, F., and H. Forsey, 1996, "On the Use and Misuse of Downside Risk", *Journal of Portfolio Management* 22, pp. 35–42.

Stutzer, M., 2000, "A Portfolio Performance Index and Its Implications", *Financial Analysts Journal* 56, pp. 30–8.

Vose, D., 2000, *Risk Analysis* (Chichester: John Wiley & Sons).

4

Portfolio Resampling and Estimation Error

This chapter introduces estimation error as an additional problem for traditional Markowitz-based portfolio construction. In contrast with Chapter 6, which also deals with estimation error but with a more decision-theoretic foundation, this chapter presents two heuristic methods that seem to be widespread among practitioners. In Sections 4.1 and 4.2 we use Monte Carlo methods to visualise the effect of estimation error on portfolio construction. We will distinguish between estimation error (when the distribution parameters are stable but we do not have enough data) and non-stationarity (the distribution parameters are unstable). The next four sections focus on the main topic of this chapter: resampled efficiency as a meaningful concept for portfolio construction.[1] The chapter concludes with an interpretation of investment constraints in the light of portfolio resampling.

4.1 VISUALISING ESTIMATION ERROR: PORTFOLIO RESAMPLING

The estimated parameters used in asset allocation problems – typically, point estimates of means, variances and correlations – are calculated using just one possible realisation of a return history. Even if we assume stationarity (constant mean, non-time-dependent covariances), we can only expect point estimates of risk and return inputs to equal the true distribution parameters if our sample is very large. The difference between estimated and true distribution parameters when samples are not sufficiently large is estimation error. The effect of estimation error on optimal portfolios can be captured by a Monte Carlo procedure called portfolio resampling.[2]

Portfolio resampling works like this. Suppose we estimated both the variance–covariance matrix, $\hat{\boldsymbol{\Omega}}_0$, and the return vector, $\hat{\boldsymbol{\mu}}_0$, using

T observations, where $\boldsymbol{\Omega}$ is a $k \times k$ covariance matrix of excess returns (asset return minus cash), and $\boldsymbol{\mu}$ is a $k \times 1$ vector of average excess returns. Our point estimates are random variables as they are calculated from random returns.

Now, what we need to do is to model the randomness of the inputs. Portfolio resampling does this by drawing repeatedly from the return distribution given by our point estimates. We create an equivalent statistical sample with T observations (the same as the original data). For this new, artificially created data set we can estimate new inputs $\hat{\boldsymbol{\Omega}}_1$ and $\hat{\boldsymbol{\mu}}_1$. By repeating this procedure n times we get n new sets of optimisation inputs: $\hat{\boldsymbol{\Omega}}_1, \hat{\boldsymbol{\mu}}_1$ to $\hat{\boldsymbol{\Omega}}_n, \hat{\boldsymbol{\mu}}_n$. For each of these inputs we can now calculate a new efficient frontier spanning from the minimum-variance portfolio to the maximum return portfolio. We calculate m portfolios along the frontier and save the corresponding allocation vectors $\boldsymbol{w}_{11}, \dots, \boldsymbol{w}_{1m}$ to $\boldsymbol{w}_{n1}, \dots, \boldsymbol{w}_{nm}$. Evaluating all m frontier portfolios for each of the n runs, with the original optimisation inputs $\hat{\boldsymbol{\Omega}}_0$, $\hat{\boldsymbol{\mu}}_0$, will force all portfolios to plot below the original efficient frontier. This is because any weight vector optimal for $\hat{\boldsymbol{\Omega}}_i$ and $\hat{\boldsymbol{\mu}}_i$, $i = 1, \dots, n$, cannot be optimal for $\hat{\boldsymbol{\Omega}}_0$, $\hat{\boldsymbol{\mu}}_0$. Therefore, all portfolio weights result in portfolios plotting below the efficient frontier as the weights have been derived from data that contain estimation error. Hence, the result of the resampling procedure is that estimation error in the inputs is transformed into uncertainty about the optimal allocation vector.

4.1.1 Resampling: an example

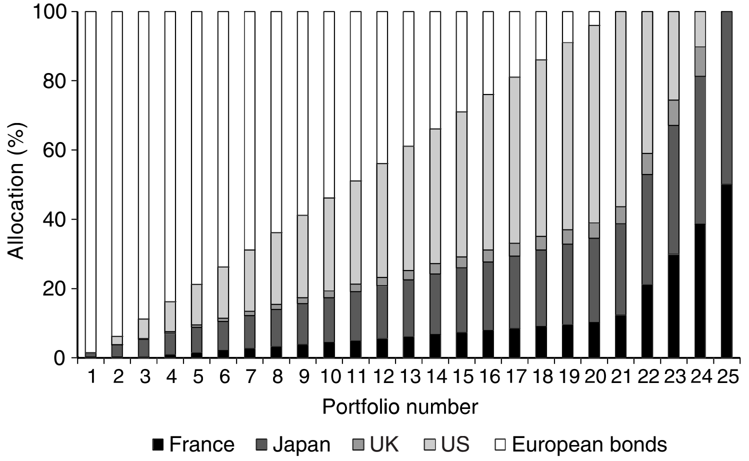
The mechanics of portfolio resampling are best illustrated by an example.[3] Suppose an analyst downloaded 18 years of data and calculated historical means and covariances, arriving at the inputs given in Table 4.1.[4]

Running a standard mean–variance optimisation (ie, minimising portfolio risk subject to a return constraint, whereby the returns vary from the return of the minimum-variance portfolio to the return of the maximum return portfolio) would, for the data set above, result in the asset allocations along the efficient frontier shown in Figure 4.1.[5] Here we have chosen to calculate $m = 25$ portfolios, dividing the difference between the minimum and maximum return into 25 steps.

Table 4.1 Input data for portfolio resampling example

	Variance–covariance matrix ($\hat{\Omega}_0$)								Return vector
Asset	**Canada**	**France**	**Germany**	**Japan**	**UK**	**US**	**US bonds**	**E bonds**	**($\hat{\mu}_0$)**
Canada	30.25	0.39							
France	15.85	49.42	0.88						
Germany	10.26	27.11	38.69	0.53					
Japan	9.68	20.79	15.33	49.56	0.88				
UK	19.17	22.82	17.94	16.92	36.12	0.79			
US	16.79	13.30	9.10	6.66	14.47	18.49	0.71		
US bonds	2.87	3.11	3.38	1.98	3.02	3.11	4.04	0.25	
Euro bonds	2.83	2.85	2.72	1.76	2.72	2.82	2.88	2.43	0.27

Figure 4.1 Mean–variance efficient portfolios

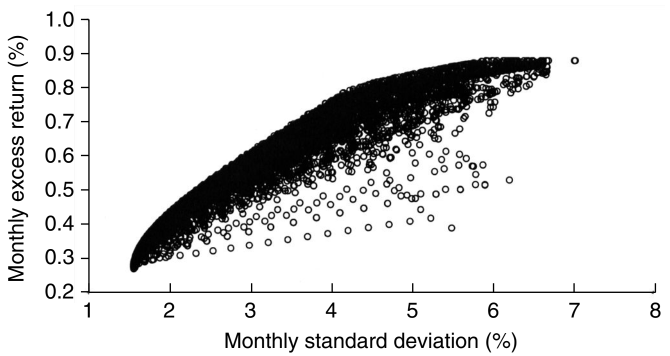


As investors familiar with a traditional portfolio optimisation would have predicted, the resulting allocations look very concentrated as some assets never even enter the solution. Also, small changes in risk aversion may lead to widely different portfolios. Allocation vectors (portfolio numbers) 20 and 23 are quite different in weightings. Given the uncertainty about the degree of investors' risk aversion, this is an unattractive feature of traditional portfolio optimisation.

Suppose instead that the resampling procedure described above is applied. Each new weight vector (calculated from resampled inputs) can be interpreted as a set of statistically equal weights. However, as only the original set of weights, w_0, is optimal for the original set of inputs, $\hat{\boldsymbol{\Omega}}_0$, $\hat{\boldsymbol{\mu}}_0$, all other portfolios must plot below the efficient frontier; their weight estimates are the direct result of sampling error. Figure 4.2 shows the efficient frontier (envelope) and the resampled portfolios derived using the resampling technique we have described. The dispersion arises owing to the great variation in statistically equal weight vectors.

Increasing the number of draws, T, forces the data points closer to the original frontier as the dispersion of inputs becomes smaller. This is equivalent to reducing sampling error. However, it is not clear from this resampling exercise where the "better" frontier lies. Section 4.3 will deal with this problem directly.

Figure 4.2 Efficient portfolios and resampled efficiency

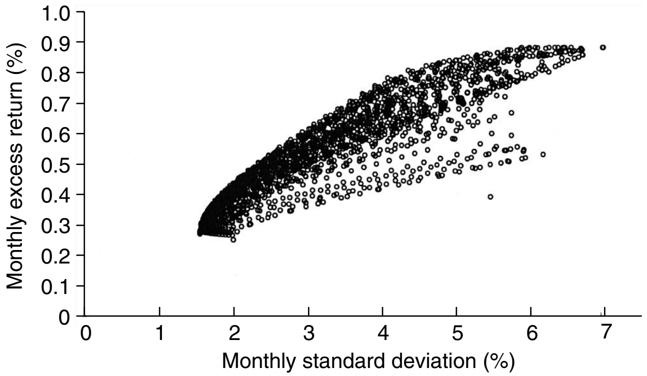


4.2 ERRORS IN MEANS AND COVARIANCES

It has long been established that portfolio optimisation suffers from error maximisation.[6] As we have seen, inputs into the efficient frontier algorithm are measured with error, and the optimiser tends to pick those assets with very attractive features (high return and low risk and/or correlation) and tends to short or deselect those with the worst features. These are exactly the cases where estimation error is likely to be highest, hence maximising the impact of estimation error on portfolio weights. If, for example, assets have high correlations, they appear similar to the quadratic programming algorithm. An algorithm that takes point estimates as inputs and treats them as if they were known with certainty (which they are not) will react to tiny differences in returns that are well within measurement error.[7] The problem gets worse as the number of assets rises because this increases the chance of outliers.

Portfolio resampling allows us to clearly distinguish the impact of the uncertainty due to estimation errors in means from that due to estimation errors in variance.[8] To measure the estimation errors in means only, we resample from the variance–covariance matrix, $\hat{\boldsymbol{\Omega}}_0$, and return vector, $\hat{\boldsymbol{\mu}}_0$, as before – the difference in the case of estimation errors in means only being that we optimise using resampled means $\hat{\boldsymbol{\mu}}_i$, $i = 1, \dots, n$, and the original variance–covariance matrix $\hat{\boldsymbol{\Omega}}_0$. The result for the data in Section 4.1 is plotted in Figure 4.3, which shows the still impressive dispersion of the risk–return points compared with Figure 4.2.

Figure 4.3 Estimation error in means only



Alternatively, we can reoptimise using resampled covariance matrixes but treating means as known. The result of this can be seen in Figure 4.4, where the data in Figure 4.3 are included for comparison. The dispersion of risk–return points is considerably reduced when estimation error is confined to variances.

So far, we have assumed that the only source of estimation error is sampling error, ie, that it is caused by insufficient data. If this were so, the problem would only be temporary – one that would not concern researchers in 200 years' time, for example, when data would be plentiful. However, there is a second source of estimation error, known as "non-stationarity". A time series is said to be non-stationary if its variance changes over time, its autocovariance is time- and also lag-dependent, or its mean changes over time.[9] When there is non-stationarity the researcher might be well advised to use shorter data sets. Extending the length of a data series might reduce the contribution of sampling error to estimation error, but at the same time it could increase that of non-stationarity.[10]

Sometimes estimation error can be observed directly from the data by estimating rolling volatilities and examining their variation. However, this approach can produce misleading results, as Figure 4.5 shows. Sampling error alone can create wide fluctuations in resampled volatilities, even for stationary data. The grey symbols in the figure are the rolling volatilities estimated from random draws from a normal distribution with the same mean and covariance as the historic data.

Figure 4.4 Estimation error in means compared with estimation error in variance

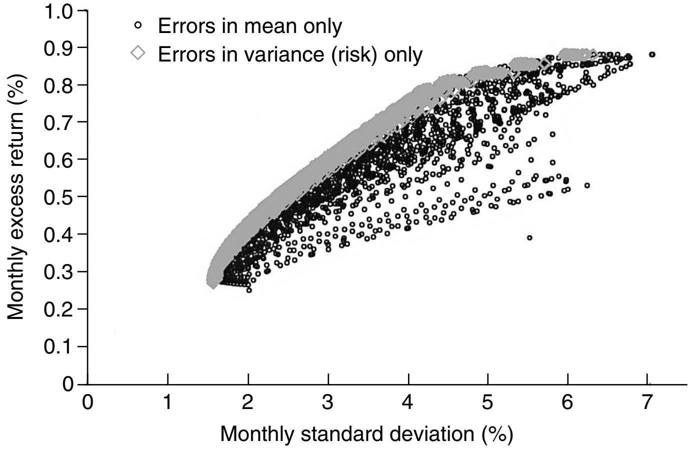
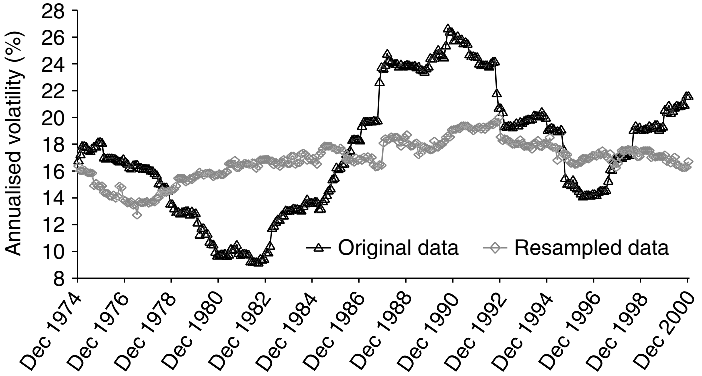


Figure 4.5 Rolling volatility of historical returns compared with rolling volatility of random draws from normal distribution with the same mean and variance



Historical data for MSCI Germany; rolling volatility calculated with 60-month window.

What we need, therefore, is a test statistic to judge whether fluctuations in rolling volatility are significant enough to reject stationarity. As we do not know the distribution of rolling volatilities under the null hypothesis, H_0, of stationarity, we have to construct a test statistic. A convenient way of doing this is Monte Carlo simulation.

Step 1. Calculate rolling volatilities for historical data using

$$\sigma_t^2(l) = \frac{1}{l}\sum_{i=t}^{t+l}(r_i - \bar{r})^2$$

where l is the window length. Record the maximum and minimum volatilities as a measure of dispersion.

Step 2. Estimate the mean and volatility of the historical series.

Step 3. Using the distribution parameters obtained in Step 2, draw from a normal distribution to create a time series of simulated returns with the same length as the historical data.[11] Record the maximum and minimum volatilities.

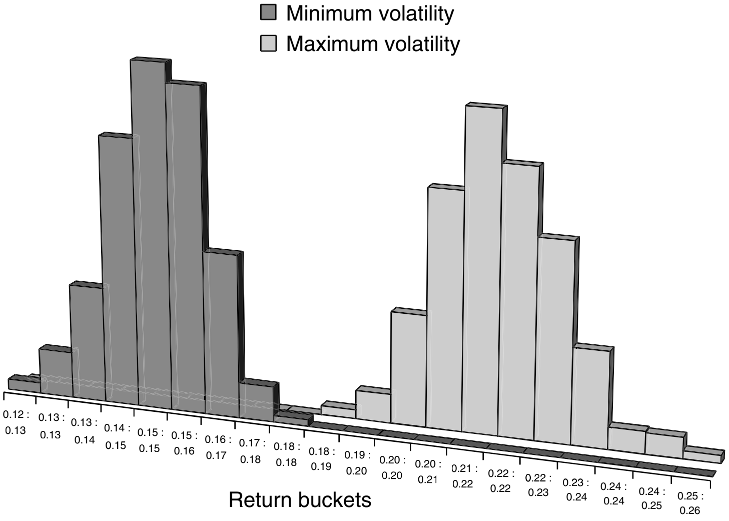
Step 4. Repeat Step 3 1,000 times. Watch the convergence of the estimates to determine the number of samplings. Then sort the results, or plot the distribution. Read the critical values for the desired confidence level from the distribution.

When this is done, we are left with the simple exercise of comparing the historical minimum and maximum volatilities (about 8% and 28%; see Figure 4.5) with their Monte Carlo derived distributions. Figure 4.6 shows the histograms for both statistics. In this case the Monte Carlo experiment supports our intuition that the historical estimates are too extreme to have been produced by sampling error – none of the 1,000 simulations produced higher maximum (or minimum) volatilities. Hence, we can conclude that it is unlikely that the data have been drawn from a stationary distribution.[12]

4.3 RESAMPLED EFFICIENCY

We have seen in Section 4.2 that the quadratic optimisation algorithm employed is too powerful for the quality of the input. This is not necessarily a problem of the mechanism itself but, rather, calls for a refinement of the input. Recently, a new concept called "resampled efficiency" has been introduced into the asset management world to deal with estimation error.[13] The objective of this and the next few sections is to describe this new technology, compare it with established procedures and point to some peculiarities of the approach.

Figure 4.6 Frequency of minimum and maximum rolling volatilities in a Monte Carlo experiment



The basis of the approach is to establish a "resampled frontier". Portfolios along this frontier are defined as an "average of the rank-associated mean–variance efficient portfolios"[14] (for further explanation see Appendix B). Averaging maintains an important portfolio characteristic – that the weights sum to one – which is probably the main practical justification for the averaging procedure. However, the method is heuristic: it has no economic justification based on the optimising behaviour of rational agents.

The resampled weight for a portfolio of rank m (portfolio number m along the frontier) is given by

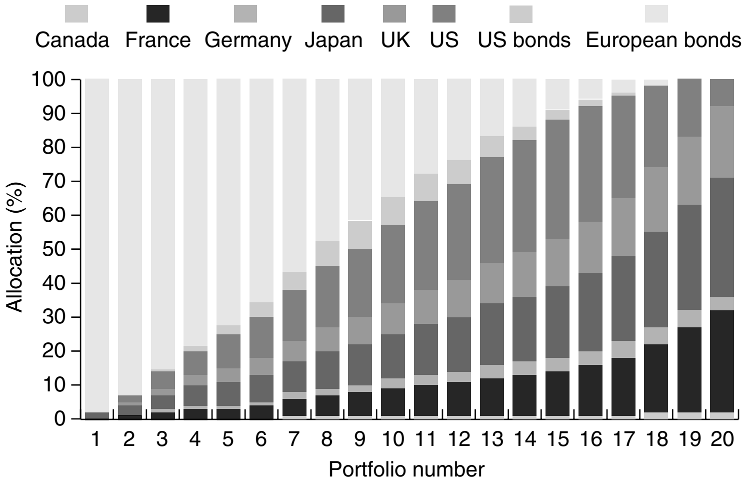
$$\bar{w}_m^{\text{resampled}} = \frac{1}{n}\sum_{i=1}^{n} w_{im} \tag{4.1}$$

where w_{im} denotes the $k \times 1$ vector of the mth portfolio along the frontier for the ith resampling. The procedure can be summarised as follows:

Step 1. Estimate the variance–covariance matrix and the mean vector of the historical inputs. (Alternatively, the inputs can be prespecified.)

Step 2. Resample, using the inputs created in Step 1, taking T draws from the input distribution; the number of draws, T, reflects

Figure 4.7 Resampled portfolios



the degree of uncertainty in the inputs. Calculate a new variance–covariance matrix from the sampled series. Estimation error will result in different matrixes from those in Step 1.

Step 3. Calculate an efficient frontier for the inputs derived in Step 2. Record the optimal portfolio weights for *m* equally distributed return points along the frontier.

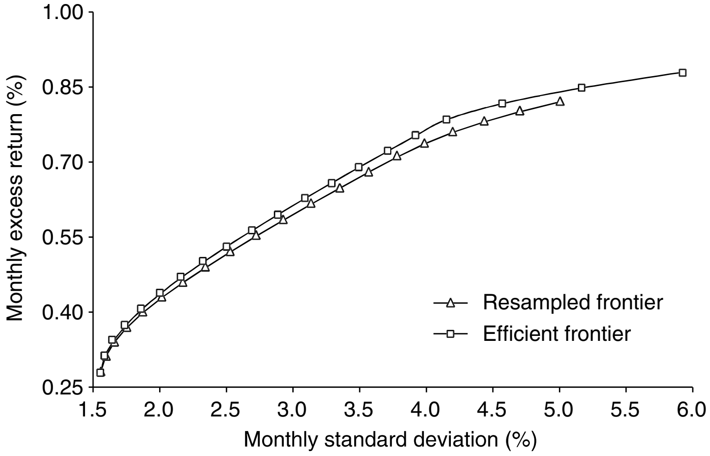
Step 4. Repeat Steps 2 and 3 many times.[15] Calculate average portfolio weights for each return point. Evaluate a frontier of averaged portfolios with the variance–covariance matrix from Step 1 to plot the resampled frontier.

Instead of adding up portfolios that share the same rank, we could alternatively add up portfolios that show the same risk–return trade-off. This can easily be done by maximising

$$U = \mu - \frac{1}{2\lambda_m}\sigma^2$$

for varying λ_m and then averaging the λ_m-associated portfolios, as demonstrated in Appendix B (on page 137). Utility-sorted portfolios are theoretically preferable as they indicate risk–return trade-offs that an investor with a given risk aversion would actually choose if in a position to repeatedly make choices under different risk–return environments.

Figure 4.8 Comparison of classic (Markowitz) frontier and resampled frontier

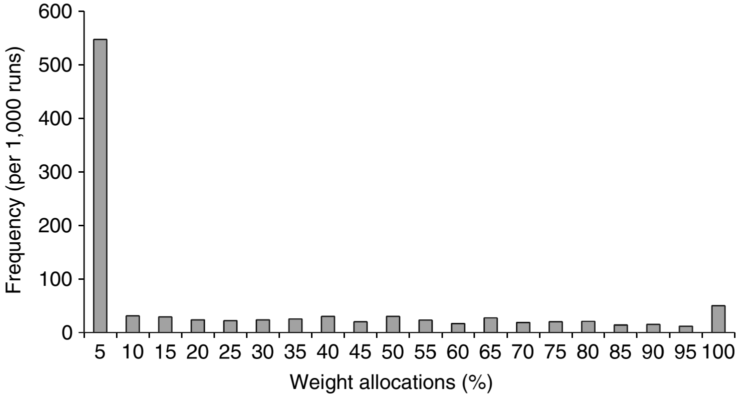


Resampled portfolios show a higher diversification, with more assets entering the solution, than classically constructed portfolios (compare Figure 4.7 with Figure 4.1). They also exhibit less sudden shifts in allocations, giving a smooth transition as return requirements change. In the eyes of many practitioners these are desirable properties.

Owing to the apparent overdiversification relative to return forecasts, the resampled frontier will show different weight allocations and will plot below the classically derived efficient frontier, as shown in Figure 4.8, and it does not reach the same maximum return as the classic frontier. Whereas the maximum return solution in the classic frontier contains only the maximum return asset(s), the averaging process prohibits this kind of solution for the resampled frontier. Both frontiers are relatively similar in risk–return space (as seen in Figure 4.8) but quite different in weight space, a feature that can be seen by comparing Figures 4.1 and 4.7.

One of the problems of using the average criterion can be appreciated by more closely inspecting the distribution of resampled weights for a particular rank-associated portfolio. Portfolio 12 in Figure 4.7 has an average allocation into US equities of about 23%. However, if we look at the distribution of resampled US equity weights for this portfolio in Figure 4.9, we find that in most of the runs (more than 500 out of 1,000), the actual weight given was

Figure 4.9 Distribution of resampled US equity weights for portfolio rank 12



between 0% and 5% and the 20%–25% columns are trivial. The average of 23% seems to be heavily influenced by the few lucky (ie, favourable) draws leading to significant allocations. This point is discussed further in Section 4.6.

4.4 DISTANCE MEASURES

Effectively, the resampling procedure provides the distribution of portfolio weights, giving us what we need to test whether two portfolios are statistically different. The difference can be measured as distance in a k-dimensional vector space, where k is the number of assets. The Euclidean measure for the distance of a vector of portfolio weights of portfolio i ($\boldsymbol{w}_i$) to portfolio p ($\boldsymbol{w}_\mathrm{p}$) is given by

$$(\boldsymbol{w}_\mathrm{p} - \boldsymbol{w}_i)'(\boldsymbol{w}_\mathrm{p} - \boldsymbol{w}_i) \tag{4.2}$$

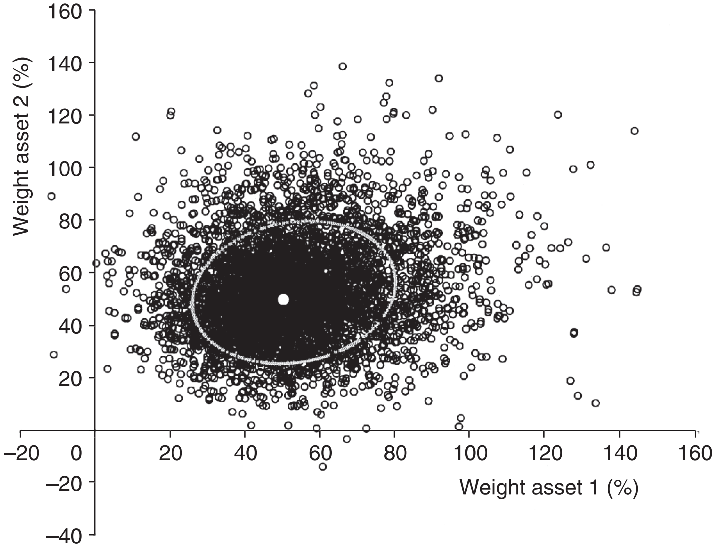
Statistical distance, however, is computed as

$$(\boldsymbol{w}_\mathrm{p} - \boldsymbol{w}_i)'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{w}_\mathrm{p} - \boldsymbol{w}_i) \tag{4.3}$$

where $\boldsymbol{\Sigma}$ is the variance–covariance matrix of portfolio weights. The test statistic represented by this expression is distributed as χ^2 with degrees of freedom equal to the number of assets, k.[16]

An example will illustrate this. Suppose we have two assets each with 10% mean and 20% volatility, that the correlation between both assets is zero and that the risk-aversion coefficient, λ, is 0.2. The

Figure 4.10 Estimation error and portfolio weights



optimal solution without estimation error is given by

$$\boldsymbol{w}^* = \begin{bmatrix} w_1^* \\ w_2^* \end{bmatrix} = \lambda \boldsymbol{\Omega}^{-1}\boldsymbol{\mu} = 0.2 \cdot \begin{bmatrix} \dfrac{1}{0.2^2} & 0 \\ 0 & \dfrac{1}{0.2^2} \end{bmatrix} \begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$

We will calculate optimal portfolios without adding-up constraints. By definition, holdings in these portfolios do not have to add up to one. In which case resampling would make no sense as all resampled weights would plot on a straight line (from 100% weight 1 to 100% weight 2). Indeed, we might be tempted to conclude that they are not portfolios as the assets do not add up, but we could think of cash as a third (filling) asset as cash would leave marginal risks as well as total risks unchanged.

Although, as we have just calculated, the optimal solution is a 50% allocation into each asset, the estimated weights are scattered around this solution. Comparing the vector difference to the critical value of χ^2 yields a measure of how statistically different a portfolio is.[17] The ellipsoid in Figure 4.10 is a line of constant density consistent with Equation 4.3 for the vector distance between the optimal portfolio without estimation error and its resamplings. In our two-asset example, lines of constant density can be obtained from

$$\left.\begin{aligned} &p(w_1 - w_1^*, w_2 - w_2^*) \\ &\quad = \frac{1}{2\pi \det(\boldsymbol{\Sigma})^{1/2}} \exp\left\{-\frac{1}{2}\begin{bmatrix} w_1 - w_1^* \\ w_2 - w_2^* \end{bmatrix}' \boldsymbol{\Sigma}^{-1} \begin{bmatrix} w_1 - w_1^* \\ w_2 - w_2^* \end{bmatrix}\right\} \\ &\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} 27.93 & 0.005 \\ 0.005 & 27.76 \end{bmatrix} \end{aligned}\right\} \tag{4.4}$$

However, introducing long-only constraints (effectively truncating weights at zero and one) invalidates the normality assumption for the distribution of portfolio weights. Michaud (1998) uses a different distance measure, widely applied in asset management, which recognises that two portfolios with the same risk and return might actually exhibit different allocations. He defines the distance between two portfolios as

$$(\boldsymbol{w}_{\mathrm{p}} - \boldsymbol{w}_i)' \hat{\boldsymbol{\Omega}}_0 (\boldsymbol{w}_{\mathrm{p}} - \boldsymbol{w}_i) \tag{4.5}$$

which is equivalent to the squared tracking error (volatility of return difference between portfolios *i* and *p*). Michaud's procedure runs as follows.

Step 1. Define a portfolio against which differences can be tested. For example, this could be the current asset allocation. Calculate Equation 4.5 for all resampled portfolios, ie, resample under new forecasts.

Step 2. Sort portfolios in descending order by tracking error, TE.

Step 3. Define TE_α as the critical tracking error for the α% level (ie, if 1,000 portfolios are resampled and the critical level is 5%, look at the tracking error of a portfolio which is 50th from the top). Hence, all portfolios for which

$$(\boldsymbol{w}_{\mathrm{p}} - \boldsymbol{w}_i)' \hat{\boldsymbol{\Omega}}_0 (\boldsymbol{w}_{\mathrm{p}} - \boldsymbol{w}_i) \geqslant \mathrm{TE}_\alpha^2 \tag{4.6}$$

are labelled statistically different. We can now ascertain whether a newly reoptimised portfolio (assuming no estimation error) is statistically different from the current asset allocation.

Step 4. Finally, calculate the minimum and maximum allocations for each asset within the confidence region.

The uncertainty about the optimal weights can be visualised for a three-asset example, but this becomes quite difficult for higher dimensions. It should be noted that similarity is defined with respect to the optimal weight vector rather than in terms of risk and return, so two portfolios could be very similar in terms of risk and return but very different in their allocations. This is a known feature as it is widely recognised that risk–return points below the frontier are not necessarily unique.

Although Michaud's test procedure is intuitive, it should be noted that the dispersion in weights is large, so it will be difficult to reject the hypothesis that both portfolios are statistically equal even if they are not. The power of the test suggested here is therefore low.

4.5 PORTFOLIO RESAMPLING AND LINEAR REGRESSION

If there were no long-only constraint, we could effectively find optimal portfolios using a simple regression approach as portfolio optimisation would then become a linear problem. Suppose we downloaded *k* (number of assets) time series of excess returns from a databank, ie, total return, *R*, minus cash rate, *c*, with *T* observations each. We can combine these excess returns in a matrix, $\boldsymbol{X}$, with each column containing one return series. Regressing these excess returns against a constant[18]

$$\mathbf{1}_{T\times 1} = \boldsymbol{X}_{T\times k}\boldsymbol{w}_{k\times 1} + \boldsymbol{u}_{T\times 1} \tag{4.7}$$

yields

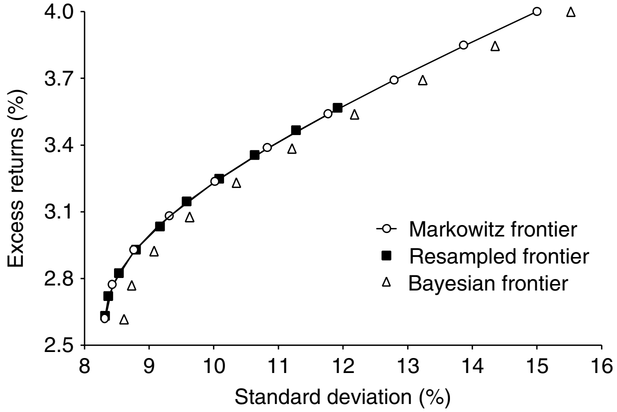
$$\hat{\boldsymbol{w}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{1} \tag{4.8}$$

where $\mathbf{1}$ denotes a $T \times 1$ vector of 1s and $\boldsymbol{u}$ is the $T \times 1$ vector of regression residuals. This can be interpreted as coming closest to a portfolio with zero risk (the vector of 1s shows no volatility) and unit return. This would present an arbitrage opportunity. Rescaling the optimal weight vector will yield the "characteristic portfolio" (one that is optimal with respect to the characteristic $\hat{\boldsymbol{\mu}}$)

$$\hat{\boldsymbol{w}}^{*}_{\text{characteristic}} = \frac{\hat{\boldsymbol{\Omega}}_0^{-1}\hat{\boldsymbol{\mu}}_0}{\boldsymbol{\mu}_0'\hat{\boldsymbol{\Omega}}_0^{-1}\hat{\boldsymbol{\mu}}_0} \tag{4.9}$$

Alternatively, a constrained regression, with a linear constraint on portfolio weights, can be run to create portfolios that meet particular return requirements. This framework can then be used to test

Figure 4.11 Bayesian frontier compared with Markowitz and resampled frontiers



restrictions on individual regression coefficients (estimated portfolio weights), as well as restrictions on groups of assets, and test whether they are significantly different from zero.[19]

Such a regression framework also puts a central problem of portfolio construction into a different, well-known perspective. Highly correlated asset returns mean highly correlated regressors, with the obvious consequences arising from multicollinearity: high standard deviations on portfolio weights (regression coefficients) and identification problems (difficulty of distinguishing between two similar assets). Simply downtesting and excluding insignificant assets will result in an outcome that is highly dependent on the order of exclusion, with no guidance on where to start.[20] This is a familiar problem both to the asset allocator and to the econometrician.

Portfolio resampling can be interpreted as a simulation approach to obtaining the distribution of weight estimates by Monte Carlo methods, as in Equation 4.7. The centre of the distribution is calculated in the same way as in portfolio resampling, ie, by averaging the coefficient estimates for a particular asset. Instead of taking the structural form of the model as given and simulating the error term, resampling simulates a whole new set of data, which is equivalent to assuming that the regressors are stochastic.[21] This can be illustrated by writing Equation 4.7 in the usual textbook form

$$y = \beta_1 x_1 + \cdots + \beta_k x_k + u, \quad \beta_i = w_i, \quad x_i = R_i - c, \quad y = 1 \quad (4.10)$$

which can be estimated using any standard regression software. Drawing new return data from the variance–covariance matrix and re-estimating Equation 4.10 n times, we can calculate the average weight for asset $j = 1, \ldots, k$ by averaging over the regression coefficients ($\hat{\beta}_i = \hat{w}_i$)

$$\bar{\hat{w}}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{w}_{ij}, \quad j = 1, \ldots, k \tag{4.11}$$

While this is not necessary for portfolios without long-only constraints (as the distribution of the regressors is known), portfolio resampling is more general than the regression approach as it can also be applied to portfolios with long-only constraints, where the weight distribution is not known. Essentially, this means bootstrapping the unknown distribution of a test statistic. If an asset is, for example, included in 70 of 1,000 runs for a given rank or utility score, it will get a p value of 7%. This approach can also be extended through Bayesian analysis using standard textbook results. Priors are set on the distribution of portfolio weights instead of asset returns.

4.6 PITFALLS OF PORTFOLIO RESAMPLING

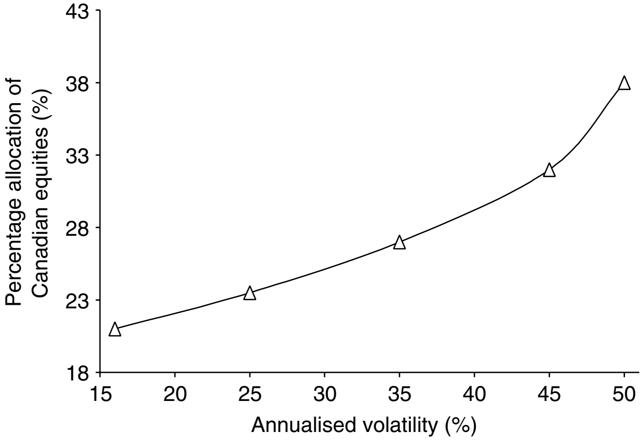
It is common sense that estimation error will result in an increase in portfolio risk. This has been captured in the Bayesian literature on portfolio construction (see Chapter 6). Taking the simplest case of a two-asset portfolio, any combination of two assets would be efficient. All resampled portfolios will still plot on the efficient frontier and no portfolio will plot below, although the frontier might be shorter as sometimes the order of assets reverses so that the averaged maximum return portfolio will not contain 100% of the higher-return asset.

Let us examine the two-asset case in more detail. Suppose that we have two uncorrelated assets with estimated volatilities of 10% and 15%, that we use 60 monthly observations to estimate the frontier and that average returns over cash are 4% and 2% per annum. Figure 4.11 plots the results for a traditional, a resampled and a Bayesian frontier.[22] The increase in risk is only captured using the Bayesian frontier, where, for the same expected return (expected returns will not change with the introduction of estimation error under uninformative priors), each portfolio is shown to expose the investor to more risk because Bayesian methods leverage up the variance–covariance matrix but leave the return vector unchanged. In direct contrast, in the resampled frontier estimation error shows up only as a shortening of the frontier, not as an increase in risk for every return level. Instead, uncertainty about the mean shows up as a reduction on the maximum expected mean return. This is not plausible.

Now let us consider two assets with the same expected return but one has significantly higher volatility. One could think of this as an international fixed-income allocation on a hedged and an unhedged basis. Most practitioners (and the mean–variance optimiser) would exclude the higher-volatility asset from the solution unless it had some desirable correlations. How would resampled efficiency deal with these assets? Repeatedly drawing from the original distribution will result in draws for the volatile asset with highly negative, as well as highly positive, returns. Quadratic programming will invest heavily in this asset in the latter case and short the asset in the former case. However, as shorting is not allowed for portfolios with long-only constraints, this will result in positive allocation for high positive average return draws and zero allocations for high negative average return draws. This is different from an unconstrained optimisation, where large long positions would be offset (on average) by large negative positions. Consequently, an increase in volatility will lead to an increase in the average allocation. Hence, a worsening Sharpe ratio would be accompanied by an increase in weight. Again, this is not a plausible result; it arises directly from the averaging rule in combination with a long-only constraint, which creates an optionality for the allocation of the corresponding asset. Assets are either in or out but never negative. This intuitive line of reasoning can be made explicit through an example based on the data set given in Table 4.1.

We will use only the equity data for Canada, France and Germany and the fixed-income data for European bonds. Also, we will reduce the sample size to 60 monthly observations, as this is a more realistic time-frame for most practical applications. We will vary only the volatility of the asset with the lowest return, Canadian equities, and look at their allocation in the maximum return portfolio (although Canadian equities have the lowest return, their allocation tends to peak at the maximum return portfolio). As volatility rises – and Sharpe ratio falls – allocations at the high-return end rise. In fact, deterioration in the risk–return relationship for Canadian equities is

Figure 4.12 Effect of volatility on allocation of Canadian equities

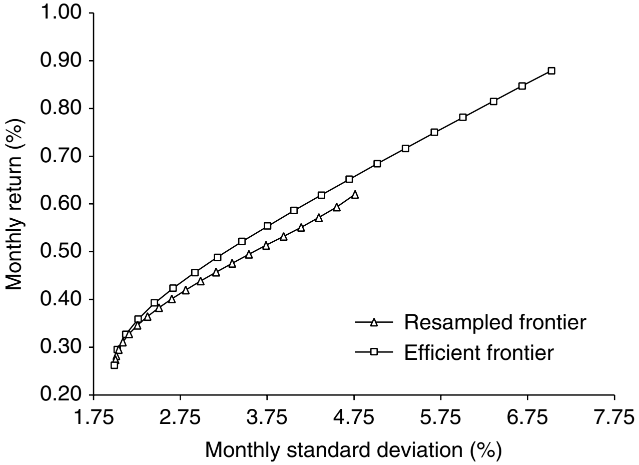


accompanied by an increased weight, as shown in Figure 4.12. This is not a result of higher volatility leading to higher estimation error – as this phenomenon does not arise in long–short portfolios – but is a direct result of averaging over long-only portfolios as the long-only constraint creates "optionality".

Another problem of resampled efficiency is that it violates basic requirements for efficient portfolios. The first such requirement is that the efficient frontier does not contain upward-bending parts.[23] Any upward-bending part would imply that we could construct portfolios superior to the frontier by linearly combining two frontier portfolios. How could such a situation arise when the concept of resampled efficiency is used? As we have already mentioned, the difference between the resampled and the classic efficient frontier is due to overdiversification.

As the return requirements rise, the maximum return portfolio tends to a state in which it contains a single asset. There might well be instances where overdiversification diminished as we moved towards the maximum return solution as these solutions do tend to be concentrated in the high-return asset. This is exactly the case with the resampled frontier shown in Figure 4.13. It is certainly true that the true test of resampled efficiency is out-of-sample performance in a Monte Carlo study. However, upward-bending parts in an efficient frontier are difficult to justify.

Figure 4.13 Upward-bending resampled frontier



A second basic violation of modern portfolio theory is that resampling (with long-only constraints) changes the structure of the maximum Sharpe ratio portfolio. This is neither intuitive nor theoretically correct as estimation error will increase the holding of cash for every level of risk aversion (as all risky assets are perceived to be riskier) but will not change the structure of the maximum Sharpe ratio portfolio. The reason for this is that risk-averse investors will increase their cash holdings as cash is not only free of volatility risk but is also free of estimation risk. Resampling, on the other hand, is very likely to change the relative weights within the maximum Sharpe ratio portfolio as it tends to overallocate to the more volatile assets. Moreover, whereas the tangency portfolio (maximum Sharpe ratio) contains no cash even in the presence of estimation error, it will always include cash in the case of resampling as cash will always be sampled in (at least for some runs).

The most important criticism of resampled efficiency arises from its statistical foundation, as all resamplings are derived from the same vector and covariance matrix, $\hat{\boldsymbol{\Omega}}_0$, $\hat{\boldsymbol{\mu}}_0$. Because the true distribution is unknown, all resampled portfolios suffer from the deviation of the parameters $\hat{\boldsymbol{\Omega}}_0$, $\hat{\boldsymbol{\mu}}_0$, from $\boldsymbol{\Omega}_{\text{true}}$, $\boldsymbol{\mu}_{\text{true}}$, in very much the same way. Averaging will not help much in this case as the averaged weights are the result of an input vector, which is itself very uncertain. Hence, it is fair to say that all portfolios inherit the same estimation error. The special importance attached to $\hat{\boldsymbol{\Omega}}_0$, $\hat{\boldsymbol{\mu}}_0$, finally limits the analysis.

4.7 CONSTRAINED PORTFOLIO OPTIMISATION

We will now leave the concept of resampled efficiency and turn to the most common procedure for avoiding "unreasonable" solutions: constrained portfolio optimisation. Constraining the set of possible solutions will, if binding, lead to a shift of the efficient frontier downwards and to the right. Although constraints are often motivated by legal constraints,[24] or preferences,[25] in this section we will assume that they are used by the investor to safeguard against "unreasonable" solutions caused by estimation error in inputs. Effectively, constraints are used to enforce some kind of diversification. Although it has already been shown that the accompanying reduction of risk is not enough to compensate for the lowered return, this section will focus on how constraints change the distribution of explicit forecasts.[26] We do this with the simulation framework used in this chapter.

Suppose that, in addition to the historical means, we have a vector of forecasts

$$\boldsymbol{\mu}_{\mathrm{f}} = \begin{pmatrix} 2\% & 1\% & 1\% & 1\% & 1\% & 1\% & 0\% & 0\% \end{pmatrix}$$

We shall also assume that the investor would prefer to avoid concentrated portfolios and so will constrain the optimal solution as given below

$$\underbrace{\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.1 \\ 0.1 \end{pmatrix}}_{w^{\text{lower}}} \leqslant \underbrace{\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}}_{\boldsymbol{A}} \underbrace{\begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \\ w_7 \\ w_8 \end{pmatrix}}_{\boldsymbol{w}} \leqslant \underbrace{\begin{pmatrix} 1 \\ 0.3 \\ 0.3 \\ 0.3 \\ 0.3 \\ 0.3 \\ 0.3 \\ 0.3 \\ 0.3 \end{pmatrix}}_{w^{\text{upper}}}$$

No asset is allowed to enter as a short position (ie, we impose a long-only constraint). No asset weight may be higher than 30% and all asset weights have to add up to 100%. Assuming λ, the risk-aversion coefficient, to be 0.038, the solution to the problem[27]

$$\left.\begin{array}{c} \boldsymbol{w}'\boldsymbol{\mu}_{\mathrm{f}} - \frac{1}{2}\lambda\boldsymbol{w}'\boldsymbol{\Omega}_0\boldsymbol{w} \\ \boldsymbol{w}^{\text{lower}} \leqslant \boldsymbol{A}\boldsymbol{w} \leqslant \boldsymbol{w}^{\text{upper}} \end{array}\right\} \tag{4.12}$$

is given by the constrained optimal allocation vector

$$\boldsymbol{w}^* = \begin{pmatrix} 30\% & 0\% & 12\% & 11\% & 0\% & 27\% & 10\% & 10\% \end{pmatrix}'$$

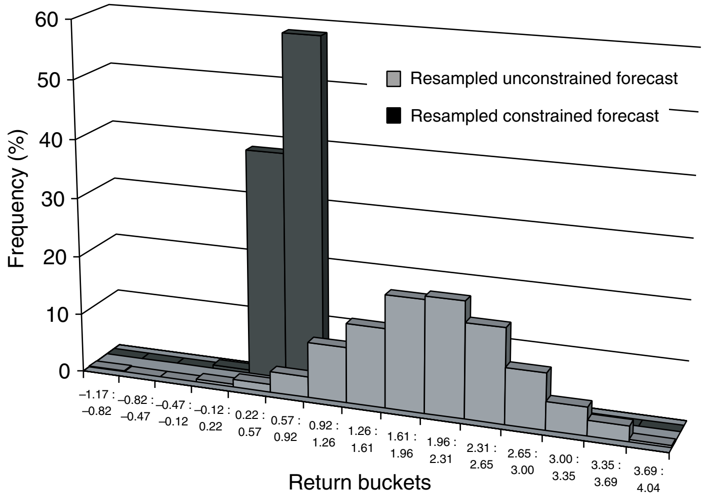
To appreciate the effect of investment constraints on optimal portfolio choice, it is informative to compare the implied returns of constrained and unconstrained portfolios. This can be done by backing out (solving for those returns that make current weights optimal) the implied returns of the new portfolio allocation to see how the constraints change our original return forecasts.[28] However, we would like to do this for a variety of optimisations, allowing us to directly compare the distribution of input return forecasts (reflecting estimation error) and the distribution of implied return forecasts from the constrained portfolios (reflecting a combination of constraints and estimation error) using the resampled portfolios. These implied returns are those that would make an investor arrive at the constrained portfolio without having constraints in the first place.

To do this we resample the optimisation in Equation 4.12 1,000 times, calculating for each run

$$\boldsymbol{\mu}_i^{\text{impl}} = \frac{\hat{\boldsymbol{\Omega}}_i \boldsymbol{w}_i^*}{\hat{\sigma}_i^{*2}} \mu_i^* = \lambda \hat{\boldsymbol{\Omega}}_i \boldsymbol{w}_i$$

The result of this procedure for asset 1 (Canadian equities) is plotted in Figure 4.14. The unconstrained implied forecasts show a much greater variation than the constrained implied forecasts. We also find that the implied returns from the constrained optimisation exhibit a significantly lower average. This is a direct effect of the binding upper constraint, which brings down the marginal contribution to risk and, hence, the implied returns. Constraining portfolio weights can be interpreted as equivalent to a high-conviction (low estimation risk) forecast. This is not necessarily intuitive as constraints have been introduced because investors do not believe in high-conviction forecasts in the first place.

Figure 4.14 Implied forecasts for Canadian equities



4.8 CONCLUSION

Portfolio resampling offers an intuitive basis for developing tests for the statistical difference between two portfolios, as represented by their weight vectors. Therefore it is generally the methodology of choice for this purpose.

It is not clear, however, why averaging over resampled portfolio weights should offer a solution for dealing with estimation error in optimal portfolio construction. In the case of long–short portfolios, averaged resampled portfolios offer no improvement over traditional Markowitz solutions; in fact, both solutions – ie, frontiers – may coincide. When long-only constraints are applied, resampled efficiency leads to more diversified portfolios; as diversified portfolios are well known to beat Markowitz portfolios out-of-sample, Michaud's (1998) finding that resampled efficiency beats simple Markowitz portfolios out-of-sample is hardly surprising.[29]

It not clear either to what extent this out-of-sample superiority can be generalised as portfolio resampling has some unwanted features, as has been shown in this chapter. Deteriorating Sharpe ratios, caused by higher volatility, lead to an increased allocation for more volatile assets in the high-return portfolios because favourable return realisations lead to large allocations, while unfavourable drawings lead to zero allocations at most (due to the phenomenon of "optionality"). Another disadvantage is that the efficient frontiers may have turning points, changing from a concave section to a convex section. It is also interesting to note that at least three assets are needed for the methodology to show the increased risk arising from estimation error.

Although the ultimate test of any portfolio construction methodology is out-of-sample performance, Markowitz methods are not the appropriate benchmark for assessing resampled efficiency; we would like to know how the latter compares with Bayesian alternatives, which do not exhibit the problems mentioned above and have a strong foundation in decision theory. Although it is not clear why resampled efficiency should be optimal, it remains an interesting heuristic for dealing with an important problem.

Constraining the set of efficient solutions is well accepted among most practitioners because it is very much in the tradition of choosing weights rather than generating return forecasts and it also expresses a deep mistrust of quantitative solutions. However, the more constrained portfolios become, the less power is given to the variance–covariance matrix and the investor ends up with an assumed, rather than a derived, solution. Resampling has shown that constraints transform a large return forecast with considerable estimation error into a smaller forecast with much less estimation error. As just stated, this is not an obvious result as the use of constraints reflects the fact that investors do not believe in high-conviction forecasts in the first place.

So far we have seen that estimates for the optimal weight vector which rely on only one source of information – historical data, or forecasts with the same uncertainty attached – are much too noisy for the optimisation algorithm. Incorporating additional information, in the form of priors, is the only way to overcome this problem. The next chapter will therefore turn to Bayesian statistics as a way of constructing more meaningful information.

APPENDIX A: LINEAR REGRESSION AND CHARACTERISTIC PORTFOLIOS

Starting from Equation 4.8 and multiplying by T/T we get

$$\hat{\boldsymbol{w}} = \left(\boldsymbol{X}'\boldsymbol{X}\frac{T}{T}\right)^{-1}\boldsymbol{X}'\mathbf{1} = \left(\boldsymbol{X}'\boldsymbol{X}\frac{1}{T}\right)^{-1}\frac{1}{T}\boldsymbol{X}'\mathbf{1} = \left(\boldsymbol{X}'\boldsymbol{X}\frac{1}{T}\right)^{-1}\hat{\boldsymbol{\mu}}_0$$

where

$$\hat{\boldsymbol{\mu}}_0 = \frac{1}{T}\boldsymbol{X}'\boldsymbol{1} = \frac{1}{T}\Big[\sum x_1 \cdots \sum x_k\Big]'$$

Rewriting the equation for the variance–covariance matrix gives

$$\begin{aligned}
\hat{\boldsymbol{\Omega}}_0 &= \frac{(\boldsymbol{X} - \boldsymbol{1}\hat{\boldsymbol{\mu}}_0')'(\boldsymbol{X} - \boldsymbol{1}\hat{\boldsymbol{\mu}}_0')}{T} \\
&= \frac{\boldsymbol{X}'\boldsymbol{X} - T\hat{\boldsymbol{\mu}}_0\hat{\boldsymbol{\mu}}_0' - T\hat{\boldsymbol{\mu}}_0\hat{\boldsymbol{\mu}}_0' - \hat{\boldsymbol{\mu}}_0\boldsymbol{1}'\boldsymbol{1}\hat{\boldsymbol{\mu}}_0'}{T} \\
&= \frac{\boldsymbol{X}'\boldsymbol{X}}{T} - \hat{\boldsymbol{\mu}}_0\hat{\boldsymbol{\mu}}_0' \\
\frac{\boldsymbol{X}'\boldsymbol{X}}{T} &= \hat{\boldsymbol{\Omega}}_0 + \hat{\boldsymbol{\mu}}_0\hat{\boldsymbol{\mu}}_0'
\end{aligned}$$

Using the result in Green (2000, p. 32, Equation 2.66), we now write

$$\begin{aligned}
\hat{\boldsymbol{w}} &= \left(\boldsymbol{X}'\boldsymbol{X}\frac{1}{T}\right)^{-1}\hat{\boldsymbol{\mu}}_0 \\
&= (\hat{\boldsymbol{\Omega}}_0 + \hat{\boldsymbol{\mu}}_0\hat{\boldsymbol{\mu}}_0')^{-1}\hat{\boldsymbol{\mu}}_0 \\
&= \left(\hat{\boldsymbol{\Omega}}_0^{-1} - \frac{\hat{\boldsymbol{\Omega}}_0^{-1}\hat{\boldsymbol{\mu}}_0\hat{\boldsymbol{\mu}}_0'\hat{\boldsymbol{\Omega}}_0^{-1}}{1 + \hat{\boldsymbol{\mu}}_0'\hat{\boldsymbol{\Omega}}_0^{-1}\hat{\boldsymbol{\mu}}_0}\right)\hat{\boldsymbol{\mu}}_0 \\
&= \frac{\hat{\boldsymbol{\Omega}}_0^{-1}\hat{\boldsymbol{\mu}}_0}{1 + \hat{\boldsymbol{\mu}}_0'\hat{\boldsymbol{\Omega}}_0^{-1}\hat{\boldsymbol{\mu}}_0}
\end{aligned}$$

Scaling this weight vector accordingly leads to the results in the text.

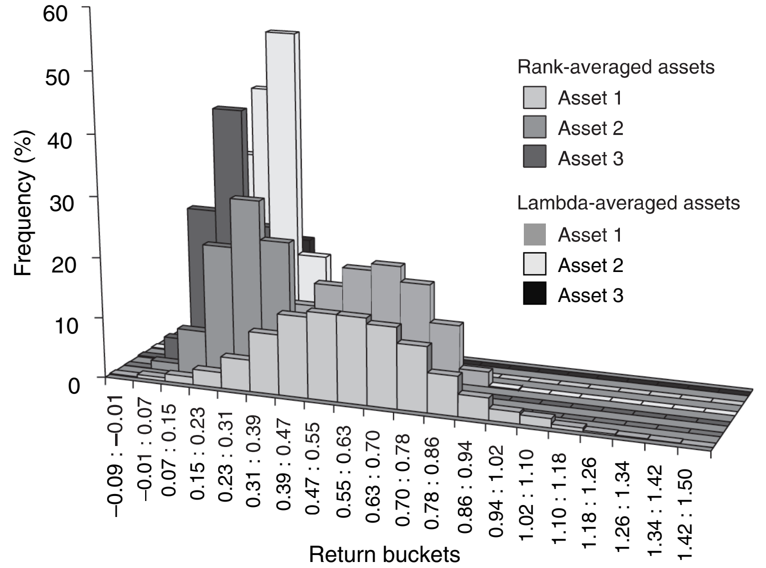
APPENDIX B: RANK-ASSOCIATED VERSUS LAMBDA-ASSOCIATED PORTFOLIOS

Resampled efficiency is defined as averaging over the rank-associated portfolios, as described in Section 4.3. Effectively, this means averaging over portfolios that have been given the same name (ie, rank $1, 2, \ldots, m$). Portfolios carrying rank 1 are the respective minimum-variance portfolios, while portfolios carrying rank m are the maximum return portfolios. All other portfolios are ranked in between according to their expected returns. The distance between the minimum-variance and maximum return portfolios is divided into equal parts.

Instead of adding up portfolios that share the same rank, we could add up portfolios that show the same risk–return trade-off. This can easily be done by maximising

$$U = \mu - \frac{1}{2\lambda_m}\sigma^2$$

Figure 4.15 Rank- and lambda-associated averages compared



for varying λ_m and then averaging the λ_m-associated portfolios. To show how close the results given by both methods are, we will resample a hypothetical three-asset case.

Assume that returns are drawn from a multivariate normal distribution with parameters

$$\boldsymbol{\mu}_0 = \begin{bmatrix} 5 \\ 10 \\ 15 \end{bmatrix}, \qquad \boldsymbol{\Omega}_0 = \begin{bmatrix} 25 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 225 \end{bmatrix}$$

We solve a simple unconstrained optimisation like $w_i^* = \lambda \boldsymbol{\Omega}_i^{-1} \boldsymbol{\mu}_i$, where i represents the ith simulation and λ has been set to 2.72. This yields the distribution of lambda-associated portfolios. Rank-associated portfolios are found by first finding out where the return of the original $(\hat{\boldsymbol{\Omega}}_0, \hat{\boldsymbol{\mu}}_0)$ lambda-associated portfolio ranks on the range between the minimum-variance and maximum return portfolios and then calculating optimal weights for each run. The distribution of resampled weights is given in Figure 4.15. Although mean allocations are virtually identical, it is interesting to note that lambda-associated portfolios show a slightly smaller variation in weights. The small experiment conducted in Figure 4.15 confirms what Michaud (1998) has already pointed out: both averaging mechanisms yield very similar results.

EXERCISES

1. Replicate Figure 4.8 in the text using S-PLUS.

2. Replicate Figure 4.9 in the text using S-PLUS.

3. Suppose you are given the following data for a three-asset portfolio optimisation exercise (calculated from 30 annual excess returns). The optimal Sharpe ratio portfolio is given by 1/6, 2/6 and 3/6

$$\bar{\boldsymbol{\Omega}}_0 = \begin{pmatrix} 400 & & \\ 210 & 225 & \\ 40 & 15 & 25 \end{pmatrix}, \qquad \bar{\boldsymbol{\mu}}_0 = \begin{pmatrix} 6.08 \\ 4.56 \\ 0.94 \end{pmatrix}$$

(a) Calculate the implied risk aversion and resample the optimisation exercise 1,000 times.

(b) Plot the distribution of portfolio weights with and without long-only constraints.

(c) Introduce a lottery ticket with zero excess returns and 40% volatility. Repeat (a) and (b) and compare the results.

4. Take a time series of hedge fund returns and calculate the maximum cumulative drawdown. Use bootstrapping to see whether this number is large or small relative to what you would expect by chance alone.

[1] The patent for the procedure, issued in December 1999 under the title "Portfolio Optimization by Means of Resampled Efficient Frontiers", is held by its inventors, Richard and Robert Michaud. The procedure is worldwide patent pending and New Frontier Advisors, LLC, Boston, MA, has been granted exclusive worldwide licensing rights.

[2] Jorion (1992) describes portfolio resampling as one method of addressing sampling error. Sampling error is the only form of estimation risk if the true underlying parameters are stable but there is insufficient data to estimate them precisely.

[3] See Scherer (2002).

[4] All examples are illustrated using the original data from Michaud (1998, pp. 17, 19). For this data $T = 216$ and $k = 8$. An entry of 30.25 stands for a standard deviation of $\sqrt{30.25} = 5.5$. Converted into annual volatility, this becomes $\sqrt{12} \times 5.5 = 19.05$. The corresponding new estimate is 0.39 per month, which is equivalent to a risk premium of 4.68 per year.

[5] Markowitz (1991) and Sharpe (1970) are classic references.

[6] See Michaud (1989) or Nawrocki (1996).

[7] This problem has been extensively reported and empirically studied. Examples are Best and Grauer (1991), Chopra and Ziemba (1993) or Jobson and Korkie (1983).

[8] It is well known that variance estimation can be improved by reducing the data interval, eg, from monthly to daily (assuming no correlation in first or second moments). The same is not true for estimates of means as the precision rises only for the average daily return and not for the annual average. A detailed exposition of this can be found in Neftci (2000).

[9] See Mills (1999, Chapter 3).

[10] Broadie (1993) is the only source (to the author's knowledge) to address non-stationarity in return series as a problem for portfolio construction.

[11] Alternatively, we could bootstrap (assigning equal probability to each return) from historical data. However, both methods yield similar results as long as the data is reasonably normal.

[12] The method described could also be used to test whether the difference between maximum and minimum volatilities (range) is significantly different from what we would expect if the data were stationary.

[13] Michaud (1998) describes his methodology well in his book *Efficient Asset Management*.

[14] Michaud (1998, p. 50).

[15] As the number of samplings grows, statistical tests can be applied with greater confidence. However, this comes at the expense of computing time. For many applications 500 samplings will suffice.

[16] The idea of this test statistic is that it is obviously not enough to look at weight differences only; small weight differences for highly correlated assets might be of greater significance than large weight differences for assets with negative correlation.

[17] Critical values of χ^2 can be found in any statistical software package, including Excel.

[18] This can be done using any econometrics package by running a regression of ones against all asset returns, excluding an intercept, hence forcing the regression through the origin in excess-return space (maximising the Sharpe ratio). See also Equation 19 in Jobson and Korkie (1983).

[19] See Britten-Jones (1999).

[20] Specifically, a procedure in which we omit the most insignificant asset, re-estimates and again omit the most insignificant asset. This is repeated until no insignificant asset is left.

[21] See Maddala (2001, p. 600) on bootstrapping data rather than residuals.

[22] See Chapter 5 on Bayesian methods with regard to portfolio construction problems.

[23] For a review of efficient set mathematics, see Huang and Litzenberger (1988).

[24] Constraints come in many forms. We can distinguish between long-only constraints (minimum holding of 0% and maximum holding of 100%), add-up constraints (portfolio weights must sum to 100%), individual constraints (upper and lower bounds on single asset), group constraints (groups of assets have to stay within boundaries), total risk constraints (portfolio beta has to stay within boundaries, volatilities must not exceed given level), risk contribution constraints (risk budgets) and return constraints (return contribution from group of assets). Implicitly we have limited the list to linear constraints.

[25] For example, a constraint with a given level of ordinary income (coupon and dividend payments) has to be maintained.

[26] See Grauer and Shen (2000).

[27] The risk-aversion parameter is derived from the risk–return trade-off implicit for an investor holding the benchmark portfolio and an unconditional return expectation for the benchmark of 3.5%.

[28] This has been suggested by Grinold and Easton (1998), who also show how to explicitly calculate which constraint contributes how much of the difference between original and implied constrained forecasts.

[29] See Jorion (1992) or Chopra *et al* (1993).

REFERENCES

Best, M., and R. Grauer, 1991, "On the Sensitivity of Mean Variance Efficient Portfolios to Changes in Asset Means: Some Analytical and Computational Results", *Review of Financial Studies* 4, pp. 314–42.

Britten-Jones, M., 1999, "The Sampling Error in Estimates of Mean Variance Efficient Portfolios", *Journal of Finance* 54, pp. 655–71.

Broadie, M., 1993, "Computing Efficient Frontiers using Estimated Parameters", *Annals of Operations Research* 45, pp. 21–58.

Chopra, V., C. Hensel and A. Turner, 1993, "Massaging Mean Variance Inputs: Returns from Alternative Global Investment Strategies, in the 1980's", *Management Science* 39, pp. 845–55.

Chopra, V., and W. Ziemba, 1993, "The Effects of Errors in Means, Variances and Covariances on Optimal Portfolio Choice", *Journal of Portfolio Management* Winter, pp. 6–11.

Grauer, R., and F. Shen, 2000, "Do Constraints Improve Portfolio Performance?", *Journal of Banking and Finance* 24, pp. 1253–74.

Green, W., 2000, *Econometric Analysis*, Fourth Edition (Englewood Cliffs, NJ: Prentice Hall).

Grinold, R., and K. Easton, 1998, "Attribution of Performance and Holdings", in W. Ziemba and J. Mulvey (eds), *Worldwide Asset and Liability Management* (Cambridge University Press).

Huang, C.-F., and R. H. Litzenberger, 1988, *Foundations for Financial Economics* (Englewood Cliffs, NJ: Prentice Hall).

Jobson, J., and B. Korkie, 1983, "Statistical Inference in Two Parameter Portfolio Theory with Multiple Regression Software", *Journal of Financial and Quantitative Analysis* 18, pp. 189–97.

Jorion, P., 1992, "Portfolio Optimization in Practice", *Financial Analysts Journal* 48, pp. 68–74.

Maddala, G., 2001, *Econometrics*, Third Edition (New York: John Wiley & Sons).

Markowitz, H., 1991, *Portfolio Selection: Efficient Diversification of Investments*, Second Edition (New York: Blackwell).

Michaud, R., 1989, "The Markowitz Optimization Enigma: Is Optimized Optimal?", *Financial Analysts Journal* 45, pp. 31–42.

Michaud, R. O., 1998, *Efficient Asset Management: A Practical Guide to Stock Portfolio Optimization and Asset Allocation* (New York: Oxford University Press).

Mills, T., 1999, *The Econometric Modelling of Financial Time Series*, Second Edition (Cambridge University Press).

Nawrocki, D., 1996, "Portfolio Analysis with a Large Universe of Assets", *Applied Economics* 28, pp. 1191–8.

Neftci, S., 2000, *An Introduction to the Mathematics of Financial Derivatives*, Second Edition (San Diego: Academic Press).

Scherer, B., 2002, "Portfolio Resampling: Review and Critique", *Financial Analysts Journal* 45, pp. 98–108.

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

5

Robust Portfolio Optimisation and Estimation Error

5.1 INTRODUCTION

Estimation error has always been acknowledged as a substantial problem in portfolio construction. Various approaches exist that range from Bayesian methods with a very strong rooting in decision theory to practitioner based heuristics with no rooting in decision theory at all as portfolio resampling. Robust optimisation is the latest attempt to address estimation error directly in the portfolio construction process. We will show that robust optimisation is equivalent to Bayesian shrinkage estimators and offer no marginal value relative to the former. The implied shrinkage that comes with robust optimisation is difficult to control. Consistent with the ad hoc treatment of uncertainty aversion in robust optimisation, we will see that out-of-sample performance largely depends on the appropriate choice of uncertainty aversion, with no guideline on how to calibrate this parameter or how to make it consistent with the more well known risk aversion. Virtually all attempts to address estimation error in portfolio construction have been around the refinement of expected returns before they enter the portfolio construction process. The error maximising property of traditional portfolio optimisation (assets with positive estimation error are over-weighted, while assets with negative estimation error are under-weighted) has been felt as a major obstacle in achieving a more scientific approach to investing. Financial economists tried to control the variation in expected returns with some form of shrinkage to either equal returns (James–Stein approach) or implied market returns (Black–Litterman approach) in the hope to also control the variation in portfolio output (and hence to arrive at less extreme and more stable solutions). Success has been mixed. First, return estimates still show outlier dependency, whatever statistical method is used. Second, parameter ambiguity will always be present, even if we increase the amount of extra sample information. But this means that error maximisation still affects portfolio construction.

Lately engineers and operations research academics have become interested in the field of portfolio optimisation and suggested two variation to mainstream thinking. The first was the idea of robust statistics which promotes the clever removal (or down-weighting) of what are thought to be extreme observations (outliers). While outliers are sometimes the only information we have got (eg, in hedge fund returns, where one manager bets against extreme events) it has been broadly felt that outlier removal reduced portfolio risk, rather than increasing it, as we would expect in the face of model error. This runs against the intuition of most portfolio and risk managers. The second addition to mainstream finance has been robust optimisation. On an intuitive level robust optimisation attempts at minimising the worst case return for a given confidence region (without confidence region the worst case return is always −100%) subject to the usual constraints. Practitioners feel this is a conservative and hence prudent form of portfolio construction, with estimation error directly built into the portfolio optimisation process. While in general this helps to dampen the error maximisation problem we will show that what looks like an innovation can be written in terms of ordinary shrinkage estimation and that the efficient set (set of optimal portfolios across an efficient frontier) remains the same.[1]

5.2 THE TÜNTÜCÜ AND KÖNIG (2004) APPROACH

Suppose investors are ambiguous about the correct variance covariance matrix or the correct mean vector in a mean variance based portfolio optimisation. Instead they have many possible candidates in mind. More precisely there exists a set of mean vectors and covariance matrixes $\boldsymbol{\mu} \in S_\mu, \boldsymbol{\Omega} \in S_\Omega$, where S_μ is the set of all mean vectors and S_Ω and is the set of all covariance matrixes. All matrixes are given equal importance no matter how unlikely they are in a probabilistic sense as it is assumed the decision maker cannot form probabilities. The optimisation problem now becomes

$$\max_{\boldsymbol{w}} \left(\min_{\boldsymbol{\mu} \in S_\mu, \boldsymbol{\Omega} \in S_\Omega} \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\mu} - \tfrac{1}{2} \lambda \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{w} \right) \tag{5.1}$$

In Equation 5.1 we want to maximise the worst case utility for all combinations of variance covariance matrixes with respect to the portfolio weight vector $\boldsymbol{w}$. The idea is to provide "good" solutions for all possible parameter realisations. We will see that this in reality means to be very pessimistic as the solution has to provide a "good" outcome even if the worst parameter specification becomes true. Problems like this can be reformulated to fit traditional optimisation software.[2] Essentially we maximise the worst case utility. For a large number of securities and a large set of mean vectors and covariance matrixes it becomes infeasible to solve Equation 5.1. However, Tüntücü and König (2004) have shown that under the assumption of a long-only constraint (all assets must be held in non-negative quantities) Equation 5.1 can be replaced by

$$\max_{\boldsymbol{w} \geqslant \boldsymbol{0}} \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\mu}_{\mathrm{l}} - \tfrac{1}{2} \lambda \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\Omega}_{\mathrm{h}} \boldsymbol{w} \tag{5.2}$$

where $\boldsymbol{\mu}_{\mathrm{l}}$ is the worst case return vector and $\boldsymbol{\Omega}_{\mathrm{h}}$ is the worst case covariance matrix. The reason we can readily identify the worst case inputs rests on the imposed long-only constraint. For a long-only position the worst case is a low expected return, so $\boldsymbol{\mu}_{\mathrm{l}}$ is the smallest element in S_{μ} while the worst case for $\boldsymbol{\Omega}_{\mathrm{h}}$ is little diversification so it must be the largest element in S_{Ω}. A high covariance for example would not be worst case for a long–short position as this implies that short positions are risk reducing (hedging). The same is true for expected returns. Low expected returns would actually be best case for a short position as there is on average less to lose. Tüntücü and König (2004) therefore use bootstrapping to elementwise construct $\boldsymbol{\mu}_{\mathrm{l}}$ and $\boldsymbol{\Omega}_{\mathrm{h}}$. For, say, 1,000 resamplings from the original inputs $\boldsymbol{\mu}_0$, $\boldsymbol{\Omega}_0$ we get 1,000 mean vectors and covariance matrixes.[3] We now look at the top left-hand element (variance for asset 1) in the variance covariance matrix and select the 5% largest entry across all 1,000 matrixes. This procedure is repeated for all elements[4] as well as for the mean vector. With respect to the later we select the lower 5% entries. Note that as $\boldsymbol{\Omega}_{\mathrm{h}} \geqslant \boldsymbol{\Omega}_0$ by construction it follows directly that $\boldsymbol{w}^{\mathrm{T}} \boldsymbol{\Omega}_{\mathrm{h}} \boldsymbol{w} - \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\Omega}_0 \boldsymbol{w} = \boldsymbol{w}^{\mathrm{T}} (\boldsymbol{\Omega}_{\mathrm{h}} - \boldsymbol{\Omega}_0) \boldsymbol{w} \geqslant \boldsymbol{0}$, ie, $\boldsymbol{\Omega}_{\mathrm{h}}$ is riskier. Also the dispersion of eigenvalues in $\boldsymbol{\Omega}_{\mathrm{h}}$ is much larger, ie, a larger fraction of variance is explained by a smaller number of factors. This should come as no surprise as the procedure to construct created high covariances mimicking the presence of a dominating market factor.

How should we evaluate this framework? The main problem in the authors' view is that the above approach translates investment risk into estimation error. Most prominently we see this with cash. Cash has neither investment nor estimation risk and in the Tüntücü and König (2004) procedure it will be the highest return asset with the lowest risk entry (zero volatility and correlation). Suppose the cash return is 2%, while a given risky asset is distributed with a 3% risk premium and 20% volatility. These numbers have been estimated with 60 monthly observations. If we identify the "worst case" expected return as a three standard deviation event the respective entry in $\boldsymbol{\mu}_\mathrm{l}$ will become $5\% - 2(20\%/\sqrt{60}) = -0.164\%$,[5] which is considerably beneath cash. For any optimisation based on Equation 5.2 with an investment universe containing both risky assets and cash will end up with a 100% cash holding as long as we look deep enough into the estimation error tail.[6] This seems to be overly pessimistic.[7] Moreover, we can see from Equation 5.2 that the Tüntücü–König formulation is equivalent to very narrow Bayesian priors. Investors would get the same result by putting a 100% weight on their priors about $\boldsymbol{\mu}_\mathrm{l}$ and $\boldsymbol{\Omega}_\mathrm{h}$. This is hardly a plausible proposition.

5.3 BOX CONSTRAINTS

One of the simplest and yet most intuitive ways to describe the uncertainty in inputs is to express the true (yet unknown) centre of the return distribution, μ_i, as part of an interval (also called a box) of length δ_i around the estimated return $\bar{\mu}_i$

$$\mu_i \in [\bar{\mu} - \delta_i, \bar{\mu} + \delta_i] \tag{5.3}$$

The framework does not constrain us to interpret $\bar{\mu}_i$ as the historical sample mean or $\pm\delta_i$ as the statistical confidence band $\delta_i = 1.96\sigma_i/\sqrt{n}$ (where σ_i represents the standard deviation of returns and n the number of available observations), although it is often used in this way. In formulating the optimisation problem we need to ensure that if an asset enters with a negative (positive) weight, the worst case expected return is $\bar{\mu}_i + \delta_i$ ($\bar{\mu}_i - \delta_i$). Short (long) positions suffer most if the true expected return is larger (smaller) than estimated. Can we formulate this within a standard quadratic program? Suppose we try

$$\max_{w, w_+, w_-} \; \boldsymbol{w}'\bar{\boldsymbol{\mu}} - \boldsymbol{\delta}'(\boldsymbol{w}_+ + \boldsymbol{w}_-) - \tfrac{1}{2}\lambda \boldsymbol{w}'\boldsymbol{\Omega}\boldsymbol{w} \tag{5.4}$$

$$\boldsymbol{w}'\mathbf{1} = 1 \tag{5.5}$$

$$\boldsymbol{w} = \boldsymbol{w}_+ - \boldsymbol{w}_-, \qquad \boldsymbol{w}_+ \geqslant 0, \quad \boldsymbol{w}_- \geqslant 0 \tag{5.6}$$

Here 5.4 describes our modified objective function, where we have added a $k \times 1$ vector $\boldsymbol{\delta}$ containing the estimation error range (for example $1.96\sigma_i/\sqrt{n}$ for each asset $i = 1, \dots, k$). Equation 5.5 includes a standard adding-up constraint and Equation 5.6 defines two non-negative variables representing absolute weights. How does this work? Imagine an asset is given a positive weight during the optimisation process. In this case the ith entry of $\boldsymbol{w}_+$ is positive while the ith entry of $\boldsymbol{w}_-$ is set to zero according to Equation 5.6. The expected return used for the optimisation will now be $w_i\bar{\mu}_i - \delta_i w_i = w_i(\bar{\mu}_i - \delta_i)$ while it will turn out to be $w_i(\bar{\mu}_i + \delta_i)$ for negative weights. Note that assets with a "wide box" (large uncertainty interval) will be penalised most (for both long and short positions) as they are simply shrunk towards zero. From a computational standpoint Equations 5.4–5.6 are very attractive as we can use widely available quadratic programming software.

5.4 A MORE GENERAL OBJECTIVE FUNCTION

Suppose we are given an m-dimensional vector of true expected returns $\boldsymbol{\mu}$, that is distributed around a mean vector, $\bar{\boldsymbol{\mu}}$, and a known covariance matrix of estimation errors, $\boldsymbol{\Sigma}$.[8] Suppose further the known variance covariance matrix of asset returns is given by the symmetric $m \times m$ matrix $\boldsymbol{\Omega}$. Note that we focus on errors in expected returns and assume the covariance matrix of asset returns to be known, such that $\boldsymbol{\Sigma} = n^{-1}\boldsymbol{\Omega}$, where n denotes the number of observations used to estimate expected returns. We will maintain this interpretation unless otherwise mentioned. It is well known from statistics that α% of the distribution of expected returns lie within an ellipsoid defined by

$$(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}) \leqslant \kappa^2_{\alpha,m} \tag{5.7}$$

with $\kappa^2_{\alpha,m} = \chi^2_m(1-\alpha)$, $\chi^2_n(1-\alpha)$ is the inverse of a chi-squared distribution with m degrees of freedom. For $\alpha = 95\%$ and $m = 8$ we can say that 95% of all expected returns lie within a statistical distance of 15.5 as defined in Equation 5.7. Moving alongside the ellipsoid covers all possible $\boldsymbol{\mu}$ that are within a provided confidence band. We can use this relationship to assess how large the difference between estimated and realised portfolio return can become, given a particular confidence region and vector of portfolio weights. Analytically we maximise the difference between expected portfolio return $\boldsymbol{w}^{\mathrm{T}}\bar{\boldsymbol{\mu}}$ and worst case statistically equivalent portfolio returns $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}$, ie, those inputs that are along the ellipsoid defined in Equation 5.7. Hence, we solve the following optimisation problem

$$L(\bar{\boldsymbol{\mu}}, \theta) = \boldsymbol{w}^{\mathrm{T}}\bar{\boldsymbol{\mu}} - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu} - \tfrac{1}{2}\theta((\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}) - \kappa_{\alpha,m}^2) \tag{5.8}$$

where θ defines the Lagrange multiplier associated with the ellipsoid constraint. Essentially, we look for the maximum distance $\boldsymbol{w}^{\mathrm{T}}\bar{\boldsymbol{\mu}} - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}$ using $\boldsymbol{\mu}$ as choice variable for any given allocation $\boldsymbol{w}$. First take derivatives of Equation 5.8 with respect to $\boldsymbol{\mu}$ and θ

$$\frac{\mathrm{d}L}{\mathrm{d}\boldsymbol{\mu}} = -\boldsymbol{w} - \theta\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}) = 0 \tag{5.9}$$

$$\frac{\mathrm{d}L}{\mathrm{d}\theta} = (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}) - \kappa_{\alpha,m}^2 = 0 \tag{5.10}$$

Solving Equation 5.9 for $(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})$ we arrive at $(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}) = -(1/\theta)\boldsymbol{\Sigma}\boldsymbol{w}$. This can then be substituted into Equation 5.10 to get us

$$\left(-\frac{1}{\theta}\boldsymbol{\Sigma}\boldsymbol{w}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(-\frac{1}{\theta}\boldsymbol{\Sigma}\boldsymbol{w}\right) - \kappa_{\alpha,m}^2 = 0$$

$$\frac{1}{\theta^2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{w} - \kappa_{\alpha,m}^2 = 0$$

$$\frac{1}{\theta^2}\sigma^2 = \kappa_{\alpha,m}^2 \tag{5.11}$$

We now solve for $1/\theta$ to substitute this back into Equation 5.9

$$\boldsymbol{\mu} = \bar{\boldsymbol{\mu}} - \left(\frac{\kappa_{\alpha,m}}{\sigma}\right)\boldsymbol{\Sigma}\boldsymbol{w} \tag{5.12}$$

Finally, we multiply both sides by $\boldsymbol{w}^{\mathrm{T}}$ to arrive at an expression for the distance between expected portfolio return $\boldsymbol{w}^{\mathrm{T}}\bar{\boldsymbol{\mu}}$ and worst case statistically equivalent portfolio returns $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}$

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu} = \boldsymbol{w}^{\mathrm{T}}\bar{\boldsymbol{\mu}} - \frac{\kappa_{\alpha,m}}{\sigma}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{w} = \boldsymbol{w}^{\mathrm{T}}\bar{\boldsymbol{\mu}} - \kappa_{\alpha,m}\sigma \tag{5.13}$$

In other words, what is the lowest value for expected portfolio returns as we move along the α%-ellipsoid? The factor $\kappa_{\alpha,m}$ can be heuristically viewed as an aversion to estimation error (uncertainty), although we have calibrated it differently above.

Robust portfolio optimisation uses $\boldsymbol{w}^{\mathrm{T}}\bar{\boldsymbol{\mu}} - \kappa_{\alpha,m}\sigma$ instead of $\boldsymbol{w}^{\mathrm{T}}\bar{\boldsymbol{\mu}}$ as an estimate for expected returns. The portfolio construction problem then becomes[9]

$$\boldsymbol{w}_{\mathrm{rob}}^{*} = \underset{\boldsymbol{w}\in\boldsymbol{C}}{\arg\max}\, \boldsymbol{w}^{\mathrm{T}}\bar{\boldsymbol{\mu}} - \kappa_{\alpha,m}\sigma - \tfrac{1}{2}\lambda\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{w} \tag{5.14}$$

instead of

$$\boldsymbol{w}^{*}_{\text{mv}} = \arg\max_{\boldsymbol{w} \in \boldsymbol{C}} \boldsymbol{w}^{\text{T}}\bar{\boldsymbol{\mu}} - \tfrac{1}{2}\lambda \boldsymbol{w}^{\text{T}}\boldsymbol{\Omega}\boldsymbol{w} \tag{5.15}$$

for Markowitz-based portfolio optimisation. Note that $\boldsymbol{w} \in \boldsymbol{C}$ serves as a shorthand for investment constraints (full investment, non-negativity, sector neutrality, beta neutrality, etc).

Robust portfolio optimisation maximises the worst case expected portfolio return for a given confidence region subject to risk return considerations. The computational difficulty with Equation 5.14 is, that it can no longer be solved using quadratic programming (as it contains a square root) if the constraint set $\boldsymbol{C}$ contains non-negativity ($\boldsymbol{w} \geqslant \boldsymbol{0}$) constraints. We need either to apply second order cone programming[10] or use an optimiser that can handle general convex expressions.

Note that if $\kappa_{\alpha,m}$ is assumed to be large, this forces the optimal solution towards assets that are relatively free from estimation error. For $\boldsymbol{\Sigma} = n^{-1}\boldsymbol{\Omega}$ estimation and investment risk move hand in hand and a larger aversion to estimation risk also reduces investment risk. The optimal portfolio invests more heavily into less risky assets. In the extreme, cash is the only asset without estimation risk, as its return is known at the beginning of the period with certainty. We will exemplify these statements in the following sections.

5.5 DERIVING OPTIMAL PORTFOLIO WEIGHTS

In this section we will look at the implicit assumptions and properties of the robust portfolio construction mechanism. This analysis is inherently in sample in nature as it does not place weight on the actual (out-of-sample) performance of constructed portfolios, but rather checks consistency with decision theory as well as evaluating the additional properties relative to established algorithms.

We start by deriving a closed-form solution for Equation 5.14 in order to better understand the mechanics of robust optimisation. For means of comparison we first state the familiar solution to the traditional portfolio optimisation problem within our context and notation. The traditional optimisation problem is given by

$$L(\boldsymbol{w}, \theta) = \boldsymbol{w}^{\text{T}}\bar{\boldsymbol{\mu}} - \tfrac{1}{2}\lambda \boldsymbol{w}^{\text{T}}\boldsymbol{\Omega}\boldsymbol{w} + \theta(\boldsymbol{w}^{\text{T}}\boldsymbol{I} - 1) \tag{5.16}$$

where θ denotes the multiplier associated with the full investment constraint ($\boldsymbol{w}^{\text{T}}\boldsymbol{I} = 1$). After taking first-order derivatives with respect to the Lagrange multiplier and the vector of portfolio weights, solving for the Lagrange multiplier and substituting this back into the derivative with respect to portfolio weights, we arrive at the familiar solution

$$w^{*}_{\mathrm{mv}} = \frac{1}{\lambda}\boldsymbol{\Omega}^{-1}\left(\bar{\boldsymbol{\mu}} - \frac{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{1}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{1}}\mathbf{1}\right) + \frac{\boldsymbol{\Omega}^{-1}\mathbf{1}}{1^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{1}} \tag{5.17}$$

The optimal Markowitz portfolio can be written as the combination of the minimum variance portfolio

$$w_{\min} = \frac{\boldsymbol{\Omega}^{-1}\mathbf{1}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{1}} \tag{5.18}$$

that is neither dependent on preferences (λ) nor expected returns $(\bar{\boldsymbol{\mu}})$ and a speculative demand

$$w_{\mathrm{spec}} = \frac{1}{\lambda}\boldsymbol{\Omega}^{-1}\left(\bar{\boldsymbol{\mu}} - \frac{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{1}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{1}}\mathbf{1}\right) \tag{5.19}$$

that depends on those factors. Note that $(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{1}/\mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{1})$ equals the return of the minimum variance portfolio. The speculative part increases if returns (opportunities) increase or risk aversion falls. This is the familiar two-fund separation.

Robust optimisation instead aims at trading off the minimum expected return for a given level of confidence against risk. Using the same notation as in Equation 5.16, this problem can be written as maximising

$$L(\boldsymbol{w}, \theta) = \boldsymbol{w}^{\mathrm{T}}\bar{\boldsymbol{\mu}} - \kappa_{\alpha,m}n^{-1/2}\sigma_{\mathrm{p}} - \tfrac{1}{2}\lambda\sigma_{\mathrm{p}}^{2} + \theta(\boldsymbol{w}^{\mathrm{T}}\mathbf{1} - 1) \tag{5.20}$$

where $\sigma_{\mathrm{p}}^{2} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{w}$, $\mathbf{1}$ is a $m \times 1$ vector of 1s and

$$n^{-1/2}\sigma_{\mathrm{p}} = (\boldsymbol{w}^{\mathrm{T}}n^{-1}\boldsymbol{\Omega}\boldsymbol{w})^{1/2} = n^{-1/2}(\sigma_{\mathrm{p}}^{2})^{1/2}$$

The first-order condition with respect to $\boldsymbol{w}$ is given as

$$\frac{\mathrm{d}L}{\mathrm{d}\boldsymbol{w}} = \bar{\boldsymbol{\mu}} - \left(\frac{n^{-1/2}\kappa_{\alpha,m} + \lambda\sigma_{\mathrm{p}}}{\sigma_{\mathrm{p}}}\right)\boldsymbol{\Omega}\boldsymbol{w} - \lambda\mathbf{1} = \mathbf{0} \tag{5.21}$$

Note that the bracketed term in Equation 5.21 is a scalar which allows us to solve for $\boldsymbol{w}$

$$\boldsymbol{w} = \left(\frac{\sigma_{\mathrm{p}}}{n^{-1/2}\kappa_{\alpha,m} + \lambda\sigma_{\mathrm{p}}}\right)\boldsymbol{\Omega}^{-1}(\bar{\boldsymbol{\mu}} - \theta\boldsymbol{I}) \tag{5.22}$$

Transpose both sides, multiply by $\mathbf{1}$ and use $\boldsymbol{w}^{\mathrm{T}}\mathbf{1} = 1$ to arrive at

$$\begin{aligned}\boldsymbol{w}^{\mathrm{T}}\mathbf{1} &= \left(\frac{\sigma_{\mathrm{p}}}{n^{-1/2}\kappa_{\alpha,m} + \lambda\sigma_{\mathrm{p}}}\right)(\bar{\boldsymbol{\mu}}^{\mathrm{T}} - \theta\mathbf{1}^{\mathrm{T}})\boldsymbol{\Omega}^{-1}\mathbf{1} \\ &= \left(\frac{\sigma_{\mathrm{p}}}{n^{-1/2}\kappa_{\alpha,m} + \lambda\sigma_{\mathrm{p}}}\right)(b - \theta a) = 1 \end{aligned} \tag{5.23}$$

where $b = \bar{\boldsymbol{\mu}}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{1}$, $a = \mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{1}$. From Equation 5.23 we can solve for θ

$$\theta = \frac{1}{a}\left(b - \frac{n^{-1/2}\kappa_{\alpha,m} + \lambda\sigma_{\mathrm{p}}}{\sigma_{\mathrm{p}}}\right) \tag{5.24}$$

Substituting Equation 5.24 into Equation 5.22 yields

$$\begin{aligned} \boldsymbol{w}^{*}_{\mathrm{rob}} &= \left(1 - \frac{n^{-1/2}\kappa_{\alpha,m}}{\lambda\sigma^{*}_{\mathrm{p}} + n^{-1/2}\kappa_{\alpha,m}}\right)\frac{1}{\lambda}\boldsymbol{\Omega}^{-1}\left(\bar{\boldsymbol{\mu}} - \frac{\bar{\boldsymbol{\mu}}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{1}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{1}}\mathbf{1}\right) + \frac{\boldsymbol{\Omega}^{-1}\mathbf{1}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{1}} \\ &= \left(1 - \frac{n^{-1/2}\kappa_{\alpha,m}}{\lambda\sigma^{*}_{\mathrm{p}} + n^{-1/2}\kappa_{\alpha,m}}\right)\boldsymbol{w}^{*}_{\mathrm{spec}} + \boldsymbol{w}_{\mathrm{min}} \end{aligned} \tag{5.25}$$

where σ^{*}_{p} denotes the standard deviation of the optimal robust portfolio.[11] For low required confidence levels ($\kappa_{\alpha,m} \to 0$) as well as for many data ($n \to \infty$) the optimal portfolio converges to a mean variance efficient (frontier) portfolio as

$$\left(1 - \frac{n^{-1/2}\kappa_{\alpha,m}}{\lambda\sigma^{*}_{\mathrm{p}} + n^{-1/2}\kappa_{\alpha,m}}\right) \to 1$$

which results in[12]

$$\boldsymbol{w}^{*}_{\mathrm{rob}} = \frac{1}{\lambda}\boldsymbol{\Omega}^{-1}\left(\bar{\boldsymbol{\mu}} - \frac{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{1}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{1}}\mathbf{1}\right) + \frac{\boldsymbol{\Omega}^{-1}\mathbf{1}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{1}} = \boldsymbol{w}^{*}_{\mathrm{mv}} \tag{5.26}$$

However, if $\kappa_{\alpha,m} \to \infty$ or if $n \to 0$, the robust portfolio converges to the minimum variance portfolio

$$\boldsymbol{w}^{*}_{\mathrm{rob}} = \frac{\boldsymbol{\Omega}^{-1}\mathbf{1}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{1}} = \boldsymbol{w}_{\mathrm{min}} \tag{5.27}$$

We see that the robust portfolio ranges between a mean variance efficient portfolio (with speculative investment demand) and the minimum variance portfolio (that ignores the information in return estimates).

5.6 HOW WELL IS UNCERTAINTY AVERSION ROOTED IN DECISION THEORY?

The author argues, that what matters after all for investors is the predictive distribution of future returns (as it determines an investor's expected utility) given by $p(\tilde{\boldsymbol{r}} \mid \boldsymbol{r}_{\mathrm{hist}})$, where $\tilde{\boldsymbol{r}}$ denotes the future returns yet unknown. The distribution is conditioned only by the observed data r_{hist} and not by any fixed realisation of the parameter vector $\boldsymbol{\theta}$ (covariances, means). We can express the predictive distribution as[13]

$$p(\tilde{\boldsymbol{r}} \mid \boldsymbol{r}_{\mathrm{hist}}) = \int p(\tilde{\boldsymbol{r}} \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \boldsymbol{r}_{\mathrm{hist}})\,\mathrm{d}\boldsymbol{\theta} \tag{5.28}$$

From Equation 5.28 we can easily see, that it is irrelevant where the variation in future returns comes from. It could either come from estimation error $p(\boldsymbol{\theta} \mid r_{\text{hist}})$ or from the conditional distribution of asset returns $p(\tilde{\boldsymbol{r}} \mid \boldsymbol{\theta})$. In this respect, it seems to makes little sense to differentiate between model uncertainty and risk. Both are inseparable. In other words, if investors believe returns could come from an array of different distributions with different parameters, they will use Equation 5.28 to model the predictive distribution taking account for parameter uncertainty. An investor, if shown the predictive distribution, does not care how much of it is due to parameter uncertainty and how much to investment risk.

However, it should be mentioned that the literature provides conflicting views on the above. Robust optimisation can be traced back to Knight (1921), who distinguishes (without axiomatic foundations) between aversion to risk, where objective probabilities exist to guide investment decisions and aversion to uncertainty where decision makers cannot even define probabilities. In stark contrast, Savage (1951) showed that decision makers act rationally by placing a prior on the parameter space to maximise posterior expected utility as long as they satisfy a set of axioms on coherent behaviour. In fact, individuals use all available tools to calculate subjective probabilities for expected utility maximisation (SEU).[14] This framework came under attack on behavioural grounds. Ellsberg (1961) observed ambiguity aversion in a series of experiments similar to the following.[15] An urn contains 300 balls, with 200 being a mixture of blue and green and 100 being red. Participants receive 100 euros if a random draw selects a ball from a prespecified colour. Participants are asked, whether they prefer this colour to be red or blue. Alternative participants receive 100 euros if the selected ball is not from the prespecified colour. Again do you prefer red or blue? The most frequent response is red in both cases. However, if red is preferred in the first case, the subjective probability for red must be higher than for blue. This must mean that you should prefer blue in the second experiment as the probability of not receiving blue (where you now receive money for) is higher than observing blue. A choice of red in both experiments is not coherent and therefore a violation of Savage's SEU. What do we make from this? For a start this is merely empirical evidence that some investors might behave irrationally. Dismissing SEU on these grounds is similar to dismissing stochastic calculus because many students repeatedly fail in experiments called exams. In fact scientists should help individuals to make better decisions, ie, erecting a normative framework, rather than following a more descriptive approach that tries to *ex post* rationalise the Ellsberg Paradox. For example, Gilboa and Schmeidler (1989) showed that "inventing" a decision maker following the minimax principle (under a different set of axioms) could reconcile the Ellsberg Paradox. Despite the intellectual beauty of their work a major problem remains. Minimax preferences are not in any respect superior to those already established by maximising expected utility with subjective utilities. Not only do they violate Savage's sure thing principle (if decision makers prefer x to y in all possible states of the world, then they should also prefer x to y in any particular state of the world) but they can also lead to a Dutch Book outcome, a situation where someone agrees to a set of bets that cause them to lose money with probability 1.[16] In the author's mind these are more serious consequences than the Ellsberg Paradox is for SEU.

5.7 HOW DIFFERENT IS ROBUST OPTIMISATION RELATIVE TO ALREADY EXISTING METHODS?

Let us interpret Equation 5.25. The careful reader will realise that the previous result essentially views robust optimisation as a shrinkage estimator that combines the minimum variance portfolio with a speculative investment portfolio, where the weighting factor is given by

$$\left(1 - \frac{n^{-1/2}\kappa_{\alpha,m}}{\lambda\sigma_{\mathrm{p}}^{*} + n^{-1/2}\kappa_{\alpha,m}}\right) \leqslant 1 \tag{5.29}$$

Note that the weighting factor contains σ_{p}^{*}, ie, the optimal volatility of robust portfolios, which is only known after the robust portfolio has been constructed. This makes "robust shrinkage" a very in-transparent and difficult to control process as the weighting factor is endogenous. As long as estimation error aversion is positive, this term will always be smaller than one. Robust portfolio construction will not be different from a shrinkage estimator like Jorion (1986) as it simply interpolates between the minimum variance portfolio and the maximum Sharpe ratio portfolio.

Additionally, the efficient set (the set of all solutions, ie, optimal portfolio) coincides with the mean variance efficient set. Solutions for investors with a particular risk aversion only differ to the extent lower weight is given to the speculative portfolio. An alternative way to see this is to rewrite Equation 5.14 as

$$U = \boldsymbol{w}^{\mathrm{T}}\bar{\boldsymbol{\mu}} - \tfrac{1}{2}\lambda_1\sqrt{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{w}} - \tfrac{1}{2}\lambda_2\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{w}$$

Taking first derivatives yields and solving for $\boldsymbol{w}^*$ yields

$$\boldsymbol{w}^* = \frac{1}{(\lambda_1/\alpha) + \lambda_2}\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{\mu}}$$

which is essentially equal to $\boldsymbol{w}^* = (1/\lambda^*)\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{\mu}}$, where λ^* is the pseudo-risk aversion that makes Markowitz and robust optimisation coincide. In other words, the appropriate choice of λ^* in mean variance optimisation will recover the robust optimisation result. Viewed this way robust optimisation offers nothing additional apart from decreased transparency. The return adjustments are not user specified but determined during the optimisation. Also we note an increased ambiguity in parameter choice: how can we justify our choice of $\kappa_{\alpha,m}$ and how can we make it "consistent" with risk aversion?[17]

5.8 REGULARISATION CONSTRAINTS

5.8.1 Equal weighting

Robust portfolio construction aims to limit the impact of estimation error on optimal portfolio choice within the optimisation process. Taken to the extreme, the most robust portfolio solution is the portfolio that does not change. Naive diversification with $w_i = 1/k$ exactly fits this bill. While incredibly naive, recent research by DeMiguel *et al* (2007) has shown that a $1/k$ portfolio rule is very hard to beat in an empirical context (in terms of realised Sharpe ratio or certainty equivalent return) by a wide variety of state-of-the-art Bayesian adjustments. For unknown $\boldsymbol{\mu}$ (and known $\boldsymbol{\Omega}$) the authors showed the conditions under which mean variance optimisation outperforms a naive equal weighting

$$\mathrm{SR}^2_{\mathrm{mv}} - \mathrm{SR}^2_{\mathrm{ew}} - \frac{k}{T} > 0 \tag{5.30}$$

where $\mathrm{SR}^2_{\mathrm{mv}}$ and $\mathrm{SR}^2_{\mathrm{ew}}$ are the squared Sharpe ratios of an optimal mean variance and a naive equal weighted portfolio, k is the number of assets and T the number of time periods for which data is available.[18] Two requirements need to be met to expect a higher utility from sophistication. First, the number of observations relative to the number of parameters must not be not too large (in other words the impact of sampling error should not be substantial) and, second, the Sharpe ratio difference between all assets must be small (otherwise, estimation error cannot cloud material risk return differences).[19]

5.8.2 Vector norm constraints: some theoretical results

The symptom of estimation error is uncertainty in weights while its cause is uncertainty about inputs. One of the most common approaches to "robustifying" a portfolio is simply to limit the range of admissible allocations, ie, constrain portfolio weights. Rather than constraining individual weights, vector norm constraints look at all weights collectively and constrain either the sum of the absolute weights (L1-norm) or the sum of the squared weights (L2-norm). Norm constraints focus on finding the correct portfolio weights rather than establishing the correct set of inputs (reflecting uncertainty) in a portfolio optimisation problem. In other words, they treat the symptoms rather than the cause.

DeMiguel *et al* (2008) provide us with some interesting insights. They focus on the impact of norm constraints on the minimum variance portfolio to allow for analytical results. While this somewhat limits their analysis, there is renewed interest in the minimum variance portfolio with index providers (like MSCI BARRA) offering minimum variance benchmarks. Adding an L1-norm constraint

$$\|\boldsymbol{w}\|_1 = \sum_{i=1}^{k} |w_1| \leqslant \delta \tag{5.31}$$

to a standard problem of finding the minimum variance portfolio (subject to an adding-up constraint)

$$\min_{w'\mathbf{1}=1} \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{w} \tag{5.32}$$

will be equivalent to finding a short-sale constrained minimum variance portfolio. While this sounds trivial (if all absolute weights have to add up to one, there cannot be any short position, ie, no leverage), they also show that an L1-norm constraint can be rewritten as

$$-\sum_{i \in \{i: w_i < 0\}} w_i \leqslant \tfrac{1}{2}(\delta - 1) \tag{5.33}$$

In other words, the sum of all asset weights that are sold short (left-hand side) must not exceed a "short-sale budget" (right-hand side). Implementing L1-norm constraints can be done efficiently with standard portfolio optimisation software. All we need to do again is to split our portfolio weight vector in a positive and a negative part, $\boldsymbol{w}^{+} + \boldsymbol{w}^{-} = \boldsymbol{w}$, $\boldsymbol{w}^{+}, \boldsymbol{w}^{-} \geqslant 0$, and add this together with $\mathbf{1}'\boldsymbol{w}^{-} \leqslant \frac{1}{2}(\delta - 1)$ to 5.32. Alternatively, we add an L2-norm

$$\|\boldsymbol{w}\|_2 = \left(\sum_{i=1}^{k} w_1^2\right)^{1/2} \leqslant \delta \tag{5.34}$$

to our optimisation problem 5.32. Note that the previously mentioned equal weighted portfolio is a special case. For $\delta = 1/\sqrt{k}$ we arrive at $w_i = 1/k$ for all assets.

Let now us introduce a numerical example to confirm the above analysis and to allow readers to replicate the results. We study annual currency excess returns (percentage change in exchange rate plus local cash rate minus USD cash rate) for the Australian dollar (AUD), euro (EUR), sterling (GBP) and yen (YEN) as below. The hypothetical variance–covariance matrix and mean vector of currency surprises are

$$\boldsymbol{\Omega} = \begin{bmatrix} 0.017 & 0.008 & 0.005 & 0.002 \\ 0.008 & 0.010 & 0.005 & 0.003 \\ 0.005 & 0.005 & 0.007 & 0.001 \\ 0.002 & 0.003 & 0.001 & 0.012 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\mu} = \begin{bmatrix} 0.01 \\ 0.02 \\ -0.01 \\ 0.02 \end{bmatrix} \tag{5.35}$$

All estimates are against the US dollar for AUD, EUR, GBP and YEN (in that order).

While currency excess returns (also called currency surprises, ie, deviations from uncovered interest parity), should be on average around zero, they are not in our sample. Currency volatility is on average about 10%, with the AUD at the high end with an annualised volatility of about 13% ($0.13 = \sqrt{0.017}$). First, we calculate the minimum variance portfolio by solving 5.32 alone. We get

$$\boldsymbol{w}'_{\text{mv}} = \begin{bmatrix} 4.67\% & 7.21\% & 56.39\% & 31.73 \end{bmatrix} \tag{5.36}$$

Now let us combine 5.31 and 5.32 for $\delta = 1$. We arrive at the identical solution, which is what we would have expected

$$\boldsymbol{w}'_{\text{L1}} = \begin{bmatrix} 4.67\% & 7.21\% & 56.39\% & 31.73 \end{bmatrix} \tag{5.37}$$

Essentially, the L1-norm imposes a short-sale constraint. Given that the minimum variance portfolio is a long-only portfolio, this constraint does not restrict the solution and we get 5.36. Finally, we combine 5.32 and 5.34 for $\delta = 1/\sqrt{4}$ to arrive at

$$\boldsymbol{w}'_{\mathrm{L2}} = \begin{bmatrix} 25\% & 25\% & 25\% & 25\% \end{bmatrix} \tag{5.38}$$

Our numerical results confirm the theoretical work by DeMiguel *et al* (2008). However, what is more interesting is to investigate the out-of-sample properties of vector norm constraints with a more realistic risk return objective.

5.8.3 Ridge regression, Sharpe ratio and robust portfolio construction

We have seen in chapter 4 that the (unconstrained) maximum Sharpe ratio portfolio can be found from a linear regression of asset returns against a vector of 1s. This made portfolio construction extremely convenient to solve with standard regression technology

$$\boldsymbol{w}_{\mathrm{OLS}} = \underset{w_i}{\arg\min} \left[\sum_{t=1}^{n} \left(1_t - \sum_{i=1}^{k} r_{ti} w_i \right)^2 \right] \tag{5.39}$$

where r_{it} denotes the excess return of asset i at time t. The intuition of Equation 5.39 is that we want to find a portfolio of k assets that comes closest to a pure arbitrage portfolio, ie, a portfolio with payout one and zero variance. The previous section introduced L2-norm constraints on portfolio weights in an unconstrained mean variance problem. Can we combine both methods? The answer is yes. All we need do is run a so-called Ridge regression,[20] which aims at minimising a penalised residual risk expression

$$\boldsymbol{w}_{\mathrm{Ridge}} = \underset{w_i}{\arg\min} \left[\sum_{t=1}^{n} \left(1_t - \sum_{i=1}^{k} r_{ti} w_i \right)^2 + \lambda \sum_{i=1}^{k} w_i^2 \right] \tag{5.40}$$

Note that Equation 5.40 is equivalent to

$$\begin{aligned} &\arg\min \left[\sum_{t=1}^{n} \left(1_t - \sum_{i=1}^{k} r_{ti} w_i \right)^2 \right] \\ &\text{s.t.} \quad \sum_{i=1}^{k} w_i^2 \leqslant \delta \end{aligned} \tag{5.41}$$

Varying λ in Equation 5.40 is equivalent to varying δ in Equation 5.41. In matrix notation the objective within brackets in Equation 5.40 becomes

$$(\mathbf{1} - \boldsymbol{X}\boldsymbol{w})'(\mathbf{1} - \boldsymbol{X}\boldsymbol{w}) + \lambda \boldsymbol{w}'\boldsymbol{w} \tag{5.42}$$

where $\boldsymbol{X}$ is the $n \times k$ matrix of asset excess returns and $\mathbf{1}$ is an $n \times 1$ vector of 1s. The solution to minimising 5.42 is given by

$$\boldsymbol{w}_{\text{Ridge}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\mathbf{1} \tag{5.43}$$

which defaults to the solution we derived in Chapter 4 for $\lambda = 0$

$$\boldsymbol{w}_{\text{OLS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{1} \tag{5.44}$$

ie, where we did do not place any weight (importance) on the dispersion of weights. Ordinary least squares (OLS), ie, mean variance portfolio construction, is therefore a special case of robust portfolio construction using Ridge regression.

5.8.4 Case study on optimal currency overlays

Whatever a given technique promises in sample, the out-of-sample performance provides the crucial litmus test for any portfolio construction method. Vector norm constraints will create weights that exhibit more stability over time by construction. However, investors are more interested in out-of-sample risk-adjusted performance. We continue to use our data set in Equation 5.35 but apply it to model a currency overlay portfolio decision. In a currency overlay[21] a so-called overlay manager chooses long and short decisions dependent on their currency forecasts independently from the underlying assets. We offer two alternatives. The typical mean-variance solution in Equation 5.45 and the robust solution with an L2-norm constraint in Equation 5.46

$$\boldsymbol{w} = \arg\max \boldsymbol{w}'\boldsymbol{\mu} - \tfrac{1}{2}\lambda \boldsymbol{w}'\boldsymbol{\Omega}\boldsymbol{w} \tag{5.45}$$

$$\boldsymbol{w}_{\text{rob}} = \underset{\sqrt{\boldsymbol{w}'\boldsymbol{w}} \leqslant \delta}{\arg\max} \boldsymbol{w}'\boldsymbol{\mu} - \tfrac{1}{2}\lambda \boldsymbol{w}'\boldsymbol{\Omega}\boldsymbol{w} \tag{5.46}$$

For readers unfamiliar with currency overlays, we need to mention that we do not need an adding-up constraint ($\boldsymbol{w}'\mathbf{1} = 1$) as the numeraire currency (here USD) always ensures our active positions add up to 0%, ie, $w_{\text{USD}} = -\boldsymbol{w}^{\text{T}}\boldsymbol{I}$. In other words, if we are 20% overweight in the AUD, 30% overweight the euro, 10% underweight the

GBP and 30% overweight the yen (all against the USD), we are short 70% the USD against this basket of currencies. In our out-of-sample testing we perform the following steps.

Step 1. Simulate 120 monthly returns from the mean vector and covariance matrix in Equation 5.35.

Step 2. Calculate a new mean vector and new covariance matrix from the data in Step 1.

Step 3. Calculate optimal portfolios for Equations 5.45 and 5.46 while saving the respective weight vectors. Use $\lambda = 5$ and $\delta = 0.5$.

Step 4. Repeat Steps 1–3 a 1,000 times.

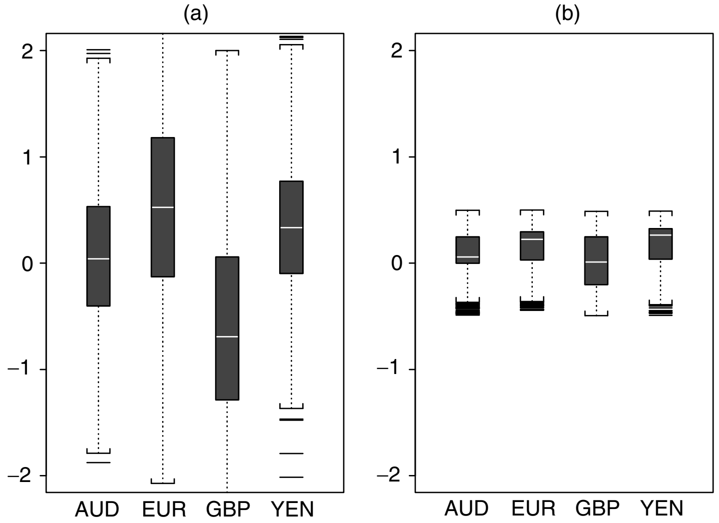
Step 5. Evaluate all portfolios (calculate utility) with the original data from Equation 5.35.

This procedure tests how valuable robustness is. For a large number of draws in Step 1, we conjecture that currency overlay management using Equation 5.45 will yield superior results, while the value of robustness is largest when the estimation error is substantial. We assume an intermediate case where our estimates come from 120 monthly data points. The distribution for all (re-)sampled weights is given in Figure 5.1. We can see that the L2-norm constraint leads to both smaller leverage (less extreme average positions) as well as a much reduced variability in optimal weights. While the Markowitz solution will chase any outlier in expected returns, the robust solution is more cautious in taking extreme positions. This is a direct consequence of the L2-norm constraint. Table 5.1 provides summary statistics for the optimal weight vectors. Again we observe that L2-norm constraints lead to less risk taking. This is only a good thing if the estimation error is large. Otherwise the investor will lose too much utility from inadequate risk taking. In other words we see that δ needs to relate to estimation error. For large estimation error, we want $\delta \to 1/\sqrt{k}$.

Unfortunately, the literature provides no guidance on how to calibrate δ as a function of risk aversion, number of observations for our estimates and number of assets in our universe.

We can now compare the difference in security equivalent between Equations 5.45 and 5.46 in Figure 5.2. For the overwhelming part of our simulations this difference is negative. On average the difference is -3.63% per annum with a t value of 31.1, ie, both economically as well as statistically significant.

Figure 5.1 The impact of a second-order norm constraint on optimal portfolio weights



We plot the optimal solutions from Equations 5.45 and 5.46 for $\delta = 0.5$ and assume that returns are drawn 120 times in each of the 1,000 simulations for $\lambda = 5$. (a) Markowitz. (b) Second-order norm constraint.

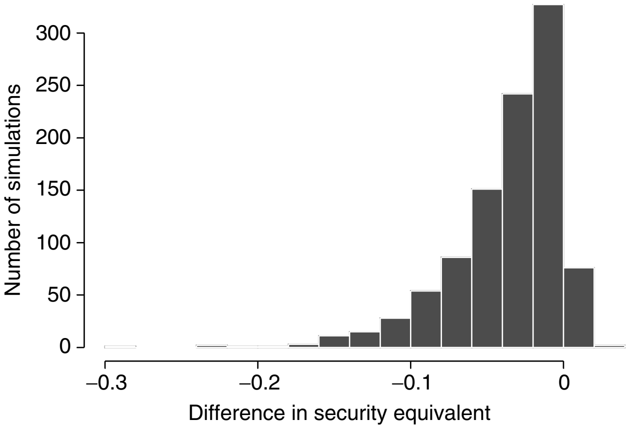
Table 5.1 Distribution of bootstrapped portfolio weights

Method	Statistics	AUD	EUR	GBP	YEN	USD
Robust (Quadratic constraint)	$\bar{w}$	9.14%	16.00%	0.44%	17.84%	−43.42%
	σ_w	20.76%	19.62%	23.83%	21.19%	45.13%
Markowitz	$\bar{w}$	6.73%	53.33%	−63.97%	33.70%	−29.78%
	σ_w	68.35%	99.50%	99.13%	65.46%	97.21%

The table provides average portfolio weights, $\bar{w}$, and their dispersion, σ_w (standard deviation) for comparison. Markowitz optimal portfolios exhibit significantly higher variation in optimal portfolio weights, approaching up to 100% (note that the ability to go short allows us to model currency weights as normally distributed).

The result so far is very supportive of using regularisation constraints. However, we need to caution the overenthusiastic users. Our results are mainly due to the fact that our investment universe is made up of assets with similar risk return ratios, ie, a situation where diversification (shrinkage) clearly pays. In the case where risk return

Figure 5.2 Difference in security equivalent



We plot a histogram for the out-of-sample difference in security equivalent between traditional Markowitz optimisation and Markowitz optimisation with a second-order norm constraint. For the overwhelming part of simulations this difference is negative. On average the difference is -3.63% per annum with a t value of 31.1, ie, both economically as well as statistically significant.

ratios are very different (eg, equities and bonds) or equivalently the number of observations is very high, experiments will show that mean-variance investing outperforms. Unfortunately, there is no clear rule how to find the optimal λ (and therefore δ) for different asset universes and sample sizes.

5.9 CONCLUSIONS

Robust portfolio optimisation aims at explicitly incorporating estimation error into the portfolio optimisation process. This chapter has formally shown that robust methods for uncertainty in mean returns are equivalent to shrinkage estimators and leave the efficient set unchanged. In other words, they offer nothing new. However, all this comes at the expense of computational difficulties (second order cone programming) and the return adjustment process is largely in-transparent relative to Bayesian alternatives.

EXERCISES

1. Create an Excel spreadsheet to implement portfolio construction.

2. Replicate the results in Section 5.8.2 using Frontline™ Risk-Solver Platform.

3. Write a program that combines Excel VBA code with the Frontline™ Risk-Solver Platform to replicate the out-of-sample results for our currency overlay example.

4. Vary the experiment in Exercise 3 by increasing the number of observations for your sample's input estimates using the code from Exercise 3.

5. Prove that Equation 5.43 is the solution to minimising 5.42.

[1] See also Scherer (forthcoming) for an out-of-sample test of robust optimisation.

[2] The mathematics has been developed by Halldorsson and Tüntücü (2003). Let us, for example, assume that there is only ambiguity about the mean vector and that the covariance matrix is known. Further assume we have 1,000 possible mean vector candidates from $\boldsymbol{\mu}_1$ to $\boldsymbol{\mu}_{1,000}$. We can reformulate Equation 5.1 into

$$\max_{w,U_{\min}} (U_{\min}), \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1 - \tfrac{1}{2}\lambda\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{w} \geqslant U_{\min}, \dots, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_{1,000} - \tfrac{1}{2}\lambda\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{w} \geqslant U_{\min}$$

NuOPT for S-Plus can deal with problems of this kind while other dedicated portfolio optimisers cannot.

[3] Note that mean vector entries are uncorrelated with covariance entries.

[4] As all covariance matrixes are symmetric, it suffices to work through the upper or lower triangle.

[5] For simplicity we assumed σ to be known.

[6] Also see Brinkmann (2005) for a review on Tüntücü and König (2004) as well as some out-of-sample tests. Using synthetic data with equal volatilities, he does not expose their method to this major deficiency and still gets only mixed results for Tüntücü–König.

[7] Maxmin criteria are known to be overly pessimistic. Their use has recently be motivated by Gilboa and Schmeidler (1989) that try to capture ambiguity by applying maxmin to expected utility. See more on this in the fourth section.

[8] This section draws on the work by Ceria and Stubbs (2005) and cleans up some of their notation. We will extent their setting in the next section.

[9] Ceria and Stubbs use the term $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{W}\|$, which is the vector norm of a product that uses the square root of a matrix. This is computationally inefficient. Using the definition of a vector norm we get

$$\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{W}\| = (\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{w})^{1/2} = (\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{w})^{1/2} = \sigma$$

which is much easier to interpret as estimation error and easier to implement.

[10] See Ghaoui *et al* (2003).

[11] Note that σ_{p}^{*} is the solution to the polynomial $a_4\sigma_{\mathrm{p}}^4 + a_3\sigma_{\mathrm{p}}^3 + a_2\sigma_{\mathrm{p}}^2 + a_1\sigma_{\mathrm{p}} + a_0$, where the coefficients depend on the above model parameters. A proof is available from the author upon request. As σ_{p}^{*} is determined endogenously, we have little control over the degree of implied shrinkage.

[12] We know that $\mathrm{d}\sigma_{\mathrm{p}}^{*}/\mathrm{d}\kappa_{\alpha,m} < 0$, ie, an increase in estimation error risk aversion will result in portfolios that carry less investment risk. This is needed to ensure that

$$1 - \frac{n^{-1/2}\kappa_{\alpha,m}}{\lambda\sigma_{\mathrm{p}}^{*} + n^{-1/2}\kappa_{\alpha,m}}$$

converges to 0 as $\kappa_{\alpha,m}$ increases.

[13] See Chapter 6 on Bayesian analysis.

[14] Given that the whole finance industry is devoting its resources to this task, this seems highly uncontroversial to the author.

[15] See Kreps (1990).

[16] See Sims (2001) for a critical view on minimax utility.

[17] Recently there has been some work on this problem. See, for example, Maenhout (2004) and the quoted literature therein. The author, however, arrives at a similar result: a dramatic decrease in the demand for risky assets.

[18] Managers of quantitative multi-strategy portfolios have known this for a long time. As long as information ratios between strategies are not too large, equal weighting works best and is hard to beat by any Bayesian adjustment mechanism.

[19] Not everybody though is willing to dismiss all information available in the data. Suppose we feel not comfortable to estimate expected returns and neither are we confident that correlation matrixes can be estimated reliably. Volatility we feel is a different matter, particularly if our asset universe is heterogeneous. In this case the "one over n" rule is replaced by the "one over volatility" rule where $w_i = 1/\sigma_i$. This can be thought of as a traditional Markowitz optimisation with a diagonal covariance matrix and forecasts that are equal to a ± 1 score multiplied by asset volatility. An extreme set of priors would obviously result in equivalent allocations.

[20] The software package S-Plus offers an extremely easy-to-use Ridge regression routine (lm.ridge) as well as a module LARS, which allows us to also use the "LASSO".

[21] See Scherer (2008) for a review of currency overlay management.

REFERENCES

Brinkmann, U., 2005, "Robuste Portfoliooptimierung: Eine kritische Bestandsaufnahme und ein Vergleich alternativer Verfahren", Internationale Tagung der Gesellschaft für Operations Research in Bremen, 09/2005.

Ceria, S., and R. Stubbs, 2005, "Incorporating Estimation Error into Portfolio Selection: Robust Efficient Frontiers", Axioma Working Paper.

DeMiguel, V., L. Garlappi and R. Uppal, 2007, "$1/N$", EFA 2006 Zurich Meetings, URL: http://ssrn.com/abstract=911512.

DeMiguel, V., L. Garlappi and R. Uppal, 2008, "A Generalized Approach to Portfolio Optimization: Improving Performance by Constraining Portfolio Norms", *Management Science* 55, pp. 798–812.

Ellsberg, D., 1961, "Risk, Ambiguity and the Savage Axioms", *Quarterly Journal of Economics* 75, pp. 643–69.

Ghaoui, L., M. Oks and F. Oustry, 2003, "Worst Case Value-at-Risk and Robust Portfolio Optimization: A Conic Programming Approach", *Operations Research* 4, pp. 543–56.

Gilboa, I., and D. Schmeidler, 1989, "Maxmin Expected Utility with Non-Unique Prior", *Journal of Mathematical Economics* 18, pp. 141–53.

Halldorsson, B., and R. H. Tüntücü, 2003, "An Interior-Point Method for a Class of Saddle Point Problems", *Journal of Optimization Theory and Applications* 116(3), pp. 559–90.

Jorion, P., 1986, "Bayes–Stein Estimation for Portfolio Analysis", *Journal of Financial and Quantitative Management* 21, pp. 279–91.

Kreps, D., 1990, *A Course in Microeconomic Theory* (Englewood Cliffs, NJ: Prentice Hall).

Knight, F. H., 1921, *Risk, Uncertainty, and Profit* (Boston, MA: Hart, Schaffner & Marx; Houghton Mifflin Company).

Maenhout, P., 2004, "Robust Portfolio Rules and Asset Pricing", *Review of Financial Studies* 17(4), pp. 951–83.

Savage, L. J., 1951, "The Theory of Statistical Decision", *Journal of the American Statistical Association* 46, pp. 55–67.

Scherer, B., 2008, "Currency Overlays", in F. J. Fabozzi (ed), *Handbook of Finance*, vol. 2, pp. 177–86 (Chichester: John Wiley & Sons).

Scherer, B., Forthcoming, "Does Robust Optimization Build Better Portfolios", *Journal of Asset Management*.

Sims, C., 2001, "Pitfalls of a Minimax Approach to Model Uncertainty", *American Economic Review* 91(2), pp. 51–4.

Tüntücü, R. H., and M. König, 2004, "Robust Asset Allocation", *Annals of Operations Research* 132, pp. 132–57.

6

Bayesian Analysis and Portfolio Choice

6.1 AN INTRODUCTION TO BAYESIAN ANALYSIS

6.1.1 Theoretical foundations

We have seen in the previous chapter that confining ourselves solely to the information available within a sample will not allow us to tackle the effect of parameter uncertainty on optimal portfolio choice. Not only do we need non-sample information (eg, additional data) to overcome this problem, but it would also be irrational of us to ignore other information based on the experience or insights – also called priors or preknowledge – of investors, statisticians and financial economists. The optimal combination of sample and non-sample information is found in Bayesian statistics. As Nobel laureate Harry Markowitz put it, "the rational investor is a Bayesian".[1]

To appreciate the implications of Bayesian statistics for portfolio choice we first need to understand the main differences between the Bayesian approach to statistics and the traditional, or "frequentist", approach. The traditional approach creates point estimates for distributional parameters. Estimates are either significant and believed to be 100% true, or insignificant and not believed at all, depending on the researcher's requirement with respect to significance.[2] For example, we either set the risk premium of currencies to zero (as we have found, statistically it has been insignificantly different from zero, but not zero itself), or we set it to a particular sample estimate that has been significantly different from zero, albeit by a narrow margin. The frequentist approach often uses maximum likelihood methods to produce point estimates, where maximising the likelihood function maximises the probability that the data have been generated from a distribution with the estimated parameters. Instead of a point estimate, Bayesian analysis produces a density function (posterior density) for the parameters involved, given the observed data. It does so by combining sample information (likelihood functions) with prior beliefs. Priors can be interpreted as the odds a researcher would be willing to accept if forced to bet on the true parameters before investigating the data. Hence, the subjective notion of probability (degree of belief) used by Bayesians is opposite to the objective notion of probability employed by frequentists (how many counts would be achieved by repeated sampling).

Suppose we have a history of risk premiums (simple returns minus cash) for a single time series of a particular asset, and that this history is summarised in a return vector $\boldsymbol{r} = (r_1\ r_2\ r_3\ \cdots\ r_T)'$, where $r_i = R_i - c_i$. Suppose also that we are interested in estimates of mean and variance summarised in a parameter vector $\boldsymbol{\theta} = (\mu\ \sigma^2)'$. Then the probability of obtaining the data (the return series) and the parameters can be written either as $p(\boldsymbol{r}, \boldsymbol{\theta}) = p(\boldsymbol{r} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$ or, alternatively, as $p(\boldsymbol{r}, \boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \boldsymbol{r})p(\boldsymbol{r})$.[3] Equating both expressions we get Bayes' Theorem

$$\underbrace{p(\boldsymbol{\theta} \mid \boldsymbol{r})}_{\substack{\text{Posterior}\\ \text{density}}} = \frac{\overbrace{p(\boldsymbol{r} \mid \boldsymbol{\theta})}^{\substack{\text{Likelihood}\\ \text{function}}}\ \overbrace{p(\boldsymbol{\theta})}^{\text{Prior}}}{p(\boldsymbol{r})} \tag{6.1}$$

The posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{r})$ of the parameters describes our information about $\boldsymbol{\theta}$ after we have observed the data, given our pre-knowledge before observing the sample information. Information in the data is captured via the likelihood function $p(\boldsymbol{r} \mid \boldsymbol{\theta})$, which is the estimated probability of observing the data if the true parameter was $\boldsymbol{\theta}$. Prior information is captured in $p(\boldsymbol{\theta})$, which gives us the odds that a risk-neutral investor would place on a bet about the value of $\boldsymbol{\theta}$. Equation 6.1 is often written in a slightly different form

$$p(\boldsymbol{\theta} \mid \boldsymbol{r}) \propto p(\boldsymbol{r} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{6.2}$$

where $\propto$ means proportional to. The reason both statements are equivalent is that $p(\boldsymbol{r})$ does not depend on $\boldsymbol{\theta}$ and, with the data already known, it can just be thought of as a normalising factor to guarantee that the area under the posterior density integrates to 1.

Equation 6.2 represents the core of Bayesian statistics. It is applied to a formulated prior distribution and the resulting expression is manipulated to find the posterior density. This is a technically complex procedure (albeit slightly less so than it might otherwise be because it is not necessary to employ maximisation techniques as in maximum likelihood estimation) and certainly beyond the scope and intention of this book, but a simple "how to do it" example is given in the next section.

In most asset allocation problems, however, we are interested in the predictive distribution (the distribution of as yet unknown future returns) for the return series as this distribution directly defines an investor's utility.[4] The predictive distribution can be written as

$$p(\tilde{\boldsymbol{r}} \mid \boldsymbol{r}) = \int p(\tilde{\boldsymbol{r}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{r}) \, \mathrm{d}\boldsymbol{\theta} \tag{6.3}$$

where $\tilde{\boldsymbol{r}}$ denotes the data not known at this time. The distribution is conditioned only by the observed data, $\boldsymbol{r}$, and not by any fixed realisation of $\boldsymbol{\theta}$. This expression looks daunting, but we can give it a simple simulation interpretation since the integral on the right-hand side can be evaluated using Monte Carlo integration.[5] First we draw n times from the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{r})$ to get n parameter vectors ($\boldsymbol{\theta}_1$ to $\boldsymbol{\theta}_n$), and then we draw from the return distribution for each parameter vector, ie, we draw from $p(\tilde{\boldsymbol{r}} \mid \boldsymbol{\theta}_1)$ to $p(\tilde{\boldsymbol{r}} \mid \boldsymbol{\theta}_n)$ to get $p(\tilde{\boldsymbol{r}} \mid \boldsymbol{r})$, where $\tilde{\boldsymbol{r}} = \begin{pmatrix} \tilde{r}_1 \, \tilde{r}_2 \, \tilde{r}_3 & \cdots & \tilde{r}_n \end{pmatrix}'$.

6.1.2 Application exercise

A simple example will illustrate the procedure.[6] Suppose that we have a single time series of risk premiums that are assumed to come from a standard Brownian motion model of asset prices. The risk premiums are normally distributed with an estimated mean, $\hat{\mu}$, of 0.3% per month and an estimated monthly standard deviation, $\hat{\sigma}$, of 6%. These estimates have been arrived at using 60 monthly observations. Suppose further that we have no prior information about the mean and variance of this series apart from what is called an "uninformed" or "uninformative" prior of the form $p(\boldsymbol{\theta}) = \sigma^{-2}$, which expresses our uncertainty about the estimates we have obtained. We want to study the effect of a lengthening time horizon in the presence of estimation error, ie, what effect estimation error has on the predictive distribution if an investor's time horizon increases.

It has been shown by Zellner (1971) that, given our assumptions, the solution to Equation 6.2, ie, the posterior density, is given by

$$p(\sigma^2 \mid \boldsymbol{r}) \sim \text{Inverse gamma}\left(\frac{T-1}{2}, \frac{1}{2}\left(\boldsymbol{r} - \frac{\boldsymbol{r}\mathbf{1}}{T}\right)'\left(\boldsymbol{r} - \frac{\boldsymbol{r}\mathbf{1}}{T}\right)\right) \tag{6.4}$$

$$p(\mu \mid \boldsymbol{r}, \sigma^2) \sim N\left(\frac{\boldsymbol{r}\mathbf{1}}{T}, \frac{\sigma^2}{T}\right) \tag{6.5}$$

We can now evaluate Equation 6.3 using a two-step procedure.

Step 1. First we draw from the posterior distribution of our parameters. To do this we first sample from Equation 6.4 1,000,000 times. We then input these 1,000,000 values of σ^2 to generate 1,000,000 values for μ. It is evident that the uninformative prior does not change the expected mean as values in Equation 6.5 are drawn from a normal distribution with a mean equal to the sample mean. This is a general feature of non-informative priors – they just add uncertainty about the range of mean returns, not about the average itself.

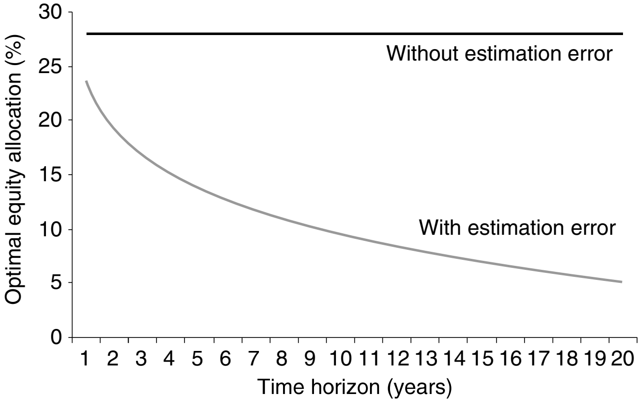
Step 2. Next we sample from our model of stock returns (a normal distribution conditional on a particular fixed parameter vector) for each pair of inputs σ^2 and μ, ie, we sample from $\tilde{r}_i \sim N(\mu_i, \sigma_i^2)$. This is repeated for different time horizons, T, by drawing from $\tilde{r}_{iT} \sim N(\mu_i T, \sigma_i^2 T)$. For each time horizon we are now able to calculate the mean and volatility of returns. We can then calculate the optimal equity allocation, as a function of time horizon, for a given risk tolerance, λ, using

$$w^*_{\text{Equity}} = \mu\lambda^{-1}\sigma^{-2} = 0.003\frac{1}{3}\frac{1}{0.06^2} = 27.7\%$$

The results for a risk tolerance of three are plotted in Figure 6.1. As the figure shows, in our simple model setting the presence of estimation error creates obvious negative time horizon effects. These are the result of an uncertainty about the mean return that adds increasingly to the total variation of returns as the horizon lengthens. After as little as one year there is already a substantial drop in the optimal equity allocation. (Appendix B (on page 199) extends this analysis by introducing forecastability.) In contrast, as shown by the upper curve in Figure 6.1, in the absence of estimation error we would get the usual result that there is no link between time horizon and the appropriate equity allocation.[7]

As we have already mentioned, prior distributions describe an investor's view before they see the data. This ensures the independence used in Equation 6.2. We can distinguish between uninformative priors and informative priors. Uninformed priors add no additional information to the data – they just express the possible range of parameter estimates. Therefore, adding an uninformative prior to our return data increases the uncertainty about future returns but

Figure 6.1 Equity allocation, time horizon and estimation error for $\lambda = 3$



not the average outcome. Informed priors do change the average expected outcome; an informed prior stating that all average returns are equal would lead us to the minimum variance portfolio, while an informed prior whose average returns were close to the return equilibrium would move us in the direction of the market portfolio.

Bayesian analysis is different from resampling (which was discussed in Chapter 4). Resampling draws repeatedly from the data without adding any new source of information to the historical data. The true distribution is still unknown (as we do not know the true parameters) and, as mentioned earlier, all resampled portfolios inherit the same estimation error relative to the true distribution.

6.1.3 Case study: a discrete prior

Suppose your task is to assess the proportion θ of active portfolio manager that can beat their benchmarks. You select a random sample of 30 active manager and find that 3 (10%) have outperformed their benchmarks. We can express this as

$$p(\text{data} \mid \theta) = \theta^3 \cdot (1-\theta)^{27}, \quad 0 < \theta < 1 \tag{6.6}$$

On the other hand your prior knowledge in Finance tells you that capital markets are efficient and as such you are opinionated, ie, you have a strong prior on how θ should be distributed. Suppose you have the prior given in Figure 6.2. With 90% probability the fraction of good managers is 1%, with 5% probability it is 5%, etc (see also Table 6.1 for the exact numbers).

Figure 6.2 Discrete prior on the fraction of good managers

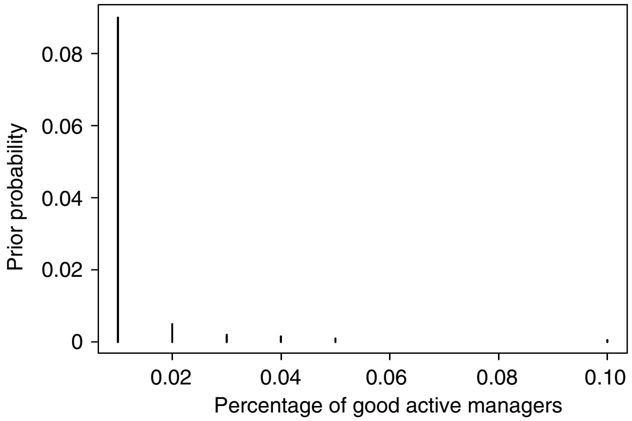


Table 6.1 Posterior distribution with discrete prior

θ (%)	$p(\theta)$ (%)	$p(\text{data} \mid \theta)$	$p(\theta)p(\text{data} \mid \theta)$	$p(\theta \mid \text{data})$ (%)
1	90.0	7.62343×10^{-7}	6.86108×10^{-7}	33
2	5.0	4.63654×10^{-6}	2.31827×10^{-7}	11
3	2.0	1.18632×10^{-5}	2.37263×10^{-7}	11
4	1.5	2.12571×10^{-5}	3.18856×10^{-7}	15
5	1.0	3.1293×10^{-5}	3.1293×10^{-7}	15
10	0.5	5.81497×10^{-5}	2.90749×10^{-7}	14
			2.07773×10^{-6}	1

Together with a discrete distribution (binomial) this is the easiest case. We will modify this example as we go along. We just need to calculate the likelihood of the data for each θ, multiply it by the prior likelihood and rescale to arrive at a proper probability distribution. This type of calculation also works for continuous distributions if you make them discrete. However, this introduces errors and is practically impossible in higher dimensions. Table 6.1 shows the involved calculations.

While you are still not convinced about the merits of active management, you are less opinionated than before. The probability of only a fraction of 1% of all managers to outperform dropped from 90% to 33% when our prior has been combined with the data.

6.1.4 Case study: a conjugate prior

We now use a conjugate prior, ie, a prior that has the same functional form as the likelihood function. The main advantage is the ability of attaining closed-form solutions. This elegance, however, comes at a price. We are restricted in our choice of prior to obey a computationally convenient form, rather than our true beliefs. Let the data come again from a binomial distribution. This time we more generically write the likelihood function as

$$p(\text{data} \mid \theta) = \theta^{y} \cdot (1-\theta)^{n-y} \tag{6.7}$$

where n denotes the number of active managers and y the number of outperforming managers. As before we assume $n = 30$, $y = 3$. Our prior is now continuous, and takes a similar form as it comes from a Beta distribution

$$p(\theta) \propto \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1} \tag{6.8}$$

where α and β are its parameters. Combining, ie, multiplying Equations 6.8 and 6.7 provides us with the posterior distribution

$$\begin{aligned} p(\text{data} \mid \theta) &\propto p(y \mid \theta)p(\theta) \\ &= \theta^{y} \cdot (1-\theta)^{n-y} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1} \\ &= \theta^{y+\alpha-1} \cdot (1-\theta)^{n-y+\beta-1} \\ &= \text{Beta}(\theta \mid \alpha+y, \beta+n-y) \end{aligned} \tag{6.9}$$

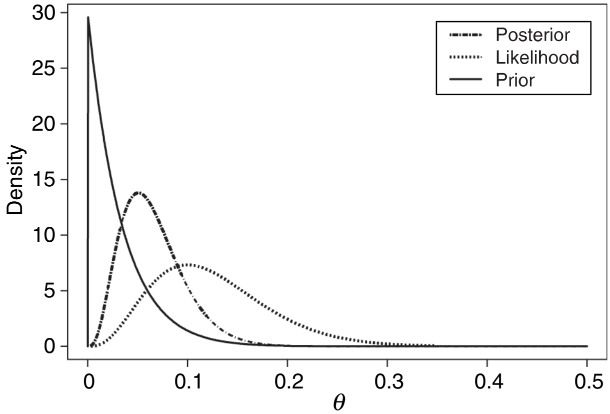
Equation 6.9 can now be used to answer questions like "what is the posterior likelihood for θ to fall between 5% and 15%?"

In Figure 6.3, we see that even when $\alpha = 1$ and $\beta = 30$ our posterior becomes considerably less "peaked" given the more favourable data.

6.2 A SIMPLE UNIVARIATE CASE

So far, we have identified estimation error as a serious problem, we have visualised the effect of uncertainty in the input on the distribution of optimal weight estimates, and we have ascertained that the addition of non-sample information is the only way to arrive at more realistic allocations. In this section we will illustrate Bayesian methods by looking at the simplest case of a single return series for which we want to estimate the mean return, assuming its variance

Figure 6.3 Posterior density with a continuous conjugate prior



is known in advance. This assumption is well justified by the small sampling error on variance estimates.

Suppose we have a time series of risk premiums

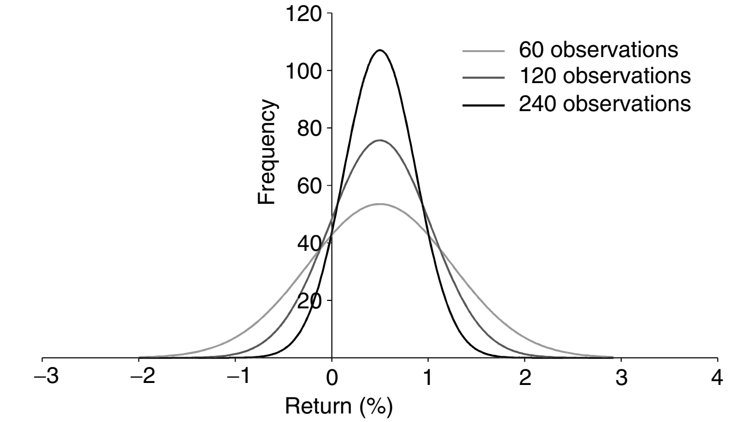
$$\boldsymbol{r} = (r_1\ r_2\ r_3\ \cdots\ r_T)'$$

for which the estimation error on the average return will depend on the number of observations available. We will estimate an average return of 0.5% per month with a monthly volatility of 5.77% (20% per annum). The same estimate carries stronger information if it is obtained using a larger sample, ie, there is less uncertainty about the average return. This relationship is plotted in Figure 6.4, from which it can be seen that, as the number of data points rises, $p(\boldsymbol{r} \mid \boldsymbol{\theta})$ becomes more "peaky" and the distribution concentrates more and more over the estimated value.[8] As the sample size approaches infinity the interval degenerates to a single point above the mean value.

How, technically, would we get the posterior distribution? We know from Equation 6.2 that $p(\mu \mid \boldsymbol{r}) \propto p(\boldsymbol{r} \mid \mu)p(\mu)$. Let us assume that our data points in $\boldsymbol{r}$ are independent draws from a normal distribution and that the conjugate prior is the exponential of a quadratic form $p(\mu) = \exp(a\mu^2 + b\mu + c)$ parameterised as[9]

$$p(\mu) \propto \exp\left(-\frac{1}{2\varphi^2}(\mu - \mu_{\text{prior}})^2\right)$$

Figure 6.4 Estimation error and sample size



Hence, we can describe $p(\mu)$ with $\mu \sim N(\mu_{\text{prior}}, \varphi^2)$. The likelihood function of T points drawn from a normal distribution with known variance is given by multiplying the likelihoods for each individual data point.[10] Equation 6.2 then becomes

$$\begin{aligned} p(\mu \mid \boldsymbol{r}) &\propto p(\mu) \prod_{i=1}^{T} p(r_i \mid \mu) \\ &\propto \exp\left(-\frac{1}{2\varphi^2}(\mu - \mu_{\text{prior}})^2\right) \prod_{i=1}^{T} \exp\left(-\frac{1}{2\sigma^2}(r_i - \mu)^2\right) \\ &\propto \exp\left\{-\frac{1}{2}\left(\frac{(\mu - \mu_{\text{prior}})^2}{\varphi^2} + \frac{\sum_{i=1}^{T}(r_i - \mu)^2}{\sigma^2}\right)\right\} \end{aligned} \tag{6.10}$$

The solution to this expression is given below (a derivation is given in Appendix A (on page 198)). The predictive mean comes from a normal distribution

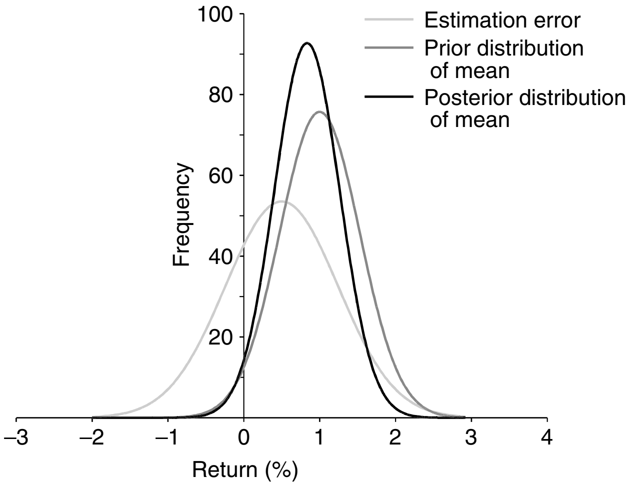
$$\mu \sim N(\mu_{\text{predictive}}, \varphi_T^2)$$

with parameters

$$\left.\begin{aligned} \mu_{\text{predictive}} &= \left(\frac{1}{\varphi_T^2}\mu_{\text{prior}} + \frac{T}{\sigma^2}\hat{\mu}\right)\left(\frac{1}{\varphi_T^2} + \frac{T}{\sigma^2}\right)^{-1} \\ \frac{1}{\varphi_T^2} &= \frac{1}{1/\varphi^2 + T/\sigma^2} \end{aligned}\right\} \tag{6.11}$$

The predictive mean, $\mu_{\text{predictive}}$, is a weighted average of the prior mean, μ_{prior}, and the historical mean, μ. The weights given to both

Figure 6.5 Effect of additional information



sources of information will depend on the precision of the respective information. Precision is defined as one divided by variance. The less volatile the information, the greater its precision. As more data becomes available the prior information becomes increasingly irrelevant and the predictive mean approaches the historical mean. The general form of Equation 6.11 can be found throughout this chapter.

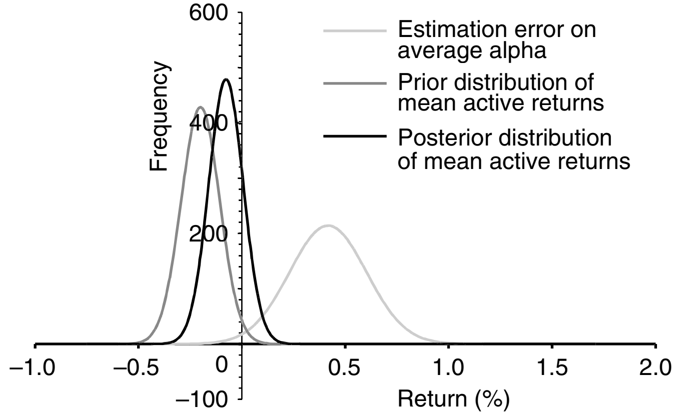
Figure 6.5 illustrates the effect of adding a second source of information when it is assumed that the prior is twice as precise as the sample estimate. When prior information is added to the data our combined estimate is more precise than either source, as can be seen from a narrower, more peaked, distribution.

We can also give a sample interpretation to Figure 6.5.[11] Suppose we have two independent samples of information and that the sample sizes differ, the second source of information containing twice as many data. If we take the grand mean (the sample mean of the combined data set) and estimate its estimation error, we arrive at the same peaked distribution for the posterior mean as in Figure 6.5.

The findings above can be summarised as follows.

- For short time series (new asset classes, initial public offerings (IPOs), new managers, etc), prior information will dominate. The longer the time series, the less important the prior will be.

Figure 6.6 Manager alpha and prior information



- Volatile asset classes exhibit large estimation errors and are therefore more amenable to the application of prior information.
- High uncertainty about the information content of a prior will result in little weight being given to this extra information.

We will finish this section with a practical application example.[12] Suppose we have historical information on an individual manager. We are confident about our risk forecast of a 5% annual tracking error, which is based on 60 monthly observations. The average annual alpha is also 5%. Additionally, we already know, from a host of empirical studies,[13] that as a direct consequence of transaction costs the average alpha in the asset management industry is likely to be negative.[14] Suppose we put a confidence in our prior assumption of −2.4% alpha per annum (equivalent to saying we are as confident as if this had been based on a sample of 240 data points), how would we view the alpha-generating capabilities of this particular manager after we combined historical data and our prior view? Substituting the inputs into Equation 6.11 and plotting the distributions yields the results presented in Figure 6.6.

Figure 6.6 shows that, in our example, prior information leads to a serious reduction in alpha expectations as the information carries little uncertainty. The posterior distribution tries to compromise between prior information and data information by finding a distribution that agrees as much as possible with both information sources. A reduced alpha forecast will obviously affect the decision whether to adopt active or passive management, as will be seen later.

6.3 A GENERAL MULTIVARIATE CASE

So far, we have introduced some basic Bayesian principles. Now we will outline a general approach that allows investors to express priors on means and variances and investigate their application to portfolio selection problems.[15] This section deals with statistical priors as we assume that informative priors are derived without reference to an economic equilibrium model. Priors derived from economic models are dealt with in the next section.[16]

Suppose we have k assets, for each of which there are T observations. Returns are assumed to follow a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Omega}$. However, their true values are unknown and we only observe their sample equivalents, ie, $\boldsymbol{r} \sim N(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Omega}})$. First we will allow priors on mean returns, assuming that the mean returns are multivariate normal as formulated below

$$\boldsymbol{\mu} \sim N(\boldsymbol{\mu}_{\text{prior}}, \boldsymbol{\Psi}_{\text{prior}}) \tag{6.12}$$

The covariance matrix $\boldsymbol{\Psi}_{\text{prior}}$ reflects our priors on the covariation of average returns, not on the covariation of asset returns. If we think, for example, that two markets should have very similar average returns (say, the Dutch and the German bond markets), we can reflect this through $\boldsymbol{\Psi}_{\text{prior}}$. We will also allow priors on the covariance of returns. The uncertainty about the true covariance matrix is modelled as an inverse Wishart distribution with v degrees of freedom[17]

$$\boldsymbol{\Omega} \sim W^{-1}(v, \boldsymbol{\Omega}_{\text{prior}}) \tag{6.13}$$

The parameter v reflects the uncertainty about our prior information on $\boldsymbol{\Omega}$, expressed via $\boldsymbol{\Omega}_{\text{prior}}$. It expresses the number of draws from $\boldsymbol{\Omega}$ that we would use to calculate the covariance matrix $\boldsymbol{\Omega}_{\text{prior}}$ if we gave our priors a sample interpretation. Setting v equal to T means that our prior information is as reliable as the historical estimate $\hat{\boldsymbol{\Omega}}$.

The predictive distribution of asset returns has both a mean vector and a covariance matrix, which can be expressed as[18]

$$\boldsymbol{\mu}_{\text{predictive}} = (\boldsymbol{\Psi}_{\text{prior}}^{-1} + \hat{\boldsymbol{\Omega}}^{-1}T)(\boldsymbol{\Psi}_{\text{prior}}^{-1}\boldsymbol{\mu}_{\text{prior}} + \hat{\boldsymbol{\Omega}}^{-1}T\hat{\boldsymbol{\mu}}) \tag{6.14}$$

$$\boldsymbol{\Omega}_{\text{predictive}} = \underbrace{\frac{T+1+v}{T+v-k-2}}_{\substack{\text{Leverage to reflect}\\ \text{estimation error}}}\left(\underbrace{\frac{v}{v+T}}_{\substack{\text{Prior}\\ \text{weight}}}\boldsymbol{\Omega}_{\text{prior}} + \underbrace{\frac{T}{v+T}}_{\substack{\text{Sample}\\ \text{weight}}}\hat{\boldsymbol{\Omega}}\right) \tag{6.15}$$

Predictive means and covariances are linear mixtures of prior and historical (sample estimate) information. Both components are weighted according to their precision.[19] In the case of equal precision, where v equals T or Ψ_{prior}^{-1} equals $\hat{\boldsymbol{\Omega}}^{-1}T$, we arrive at equal weightings. The term for the predictive variance also includes $(T+v-k-2)^{-1}(T+v+1)$, which adjusts for pure estimation risk. Additional information ($v > 0$) reduces estimation risk, whereas a larger number of assets increases estimation risk. Estimation risk enters the formula as a leverage factor on the covariance matrix. Risk is increasing as there is now uncertainty attached to point estimates, which have previously been treated as being known without uncertainty. Less data means higher covariance leverage. Uncertainty also rises with the number of assets as there is a greater likelihood that one asset will show an unusual estimate.

The framework outlined above allows us to accommodate different sets of priors. If we put ourselves in the position of having no prior opinion about the parameters (a general non-informative case, also called Jeffreys' prior), we set v equal to zero (where we have no information about covariance) and make $\boldsymbol{\Psi}_{\text{prior}}$ infinite, ie, there is a large uncertainty about the range of forecasts.[20] Equations 6.14 and 6.15 then become

$$\boldsymbol{\mu}_{\text{predictive}} = \hat{\boldsymbol{\mu}}, \qquad \boldsymbol{\Omega}_{\text{predictive}} = \frac{T+1}{T-k-2}\hat{\boldsymbol{\Omega}} \tag{6.16}$$

This confirms that if returns are multivariate normal and we assume a diffuse prior, our estimates for the mean vector remain unchanged.[21] This is intuitive as uncertainty about the expected value shows up in the difference between mean and realisation. There is, however, no uncertainty about an expected value. The variation of returns increases by a constant factor so that the new variance–covariance matrix is just leveraged up (see Equation 6.16).[22] Estimation error affects all assets in the same, proportional, way, with the scaling factor depending on sample size –

small sample sizes result in greater uncertainty and, hence, larger scalings.[23] However, instead of being multivariate normal, the predictive distribution of asset returns with estimation error follows a multivariate *t*-distribution. The admissibility of mean–variance analysis now depends critically on the assumption of quadratic utility. If investors show quadratic utility, the scaling of the covariance matrix can be interpreted as a scaling in risk aversion. In this case, estimation error will result in a higher pseudo risk aversion, and this in turn has five consequences.

- For each return requirement the optimal portfolio will show higher risk due to the scaling.
- The composition of portfolios along the efficient frontiers will stay the same because for any original risk aversion there exists an equal rescaled risk aversion.
- Investors are likely to find different portfolios as the risk–return trade-off changes, ie, flattens.
- If the utility function is not quadratic, the mean–variance representation of the problem holds only approximately; alternative models might be better suited to deal with this.[24]
- The inclusion of cash will leave the composition of the maximum Sharpe ratio portfolio unchanged. Investors will react to estimation risk with an increase in cash, which is the only asset free of investment risk and estimation risk. They will not change the composition of the minimum-variance portfolio as estimation risk affects all assets in the same way.

Rational investors will use the predictive distribution (including estimation risk and prior information) to calculate their expected utility. However, as purely non-informative priors do not change the set of efficient portfolios, such priors have been widely ignored. Instead, there has been more interest in informative priors.

How should we implement the view on volatility separately from prior information on correlations? As the covariance matrix uses correlations and volatilities simultaneously, we have to split the covariance matrix into a symmetric correlation matrix and a diagonal volatility matrix

$$\boldsymbol{\Omega}_{\text{prior}} = \underbrace{\begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_k \end{pmatrix}}_{\boldsymbol{\sigma}_{\text{prior}}}$$

$$\times \underbrace{\begin{pmatrix} 1 & \hat{\rho}_{12} & \cdots & \hat{\rho}_{1k} \\ \hat{\rho}_{21} & 1 & \cdots & \hat{\rho}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}_{k1} & \hat{\rho}_{k2} & \cdots & 1 \end{pmatrix}}_{\hat{\boldsymbol{\rho}}} \underbrace{\begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_k \end{pmatrix}}_{\boldsymbol{\sigma}_{\text{prior}}}$$

This allows us to input priors on volatilities but leave correlations unchanged. As more and more data becomes available, the prior becomes less important and we approach the sample estimates

$$\left(\boldsymbol{\mu}_{\text{predictive}} \underset{T\to\infty}{\longrightarrow} \hat{\boldsymbol{\mu}}; \boldsymbol{\Omega}_{\text{predictive}} \underset{T\to\infty}{\longrightarrow} \hat{\boldsymbol{\Omega}}\right)$$

which is a basic feature of Bayesian methods. However, this is not the case with the Black–Litterman (B–L) model, which is reviewed in the next section.

6.4 SPECIAL CASE: BLACK–LITTERMAN

6.4.1 General outline

The previous section presented a general framework for combining historical sample information with additional non-sample information. However, the problem remains that if we do not have an informative prior, we cannot arrive at an unambiguous and economically meaningful default allocation.[25] In this case it would be prudent to hold the market portfolio (combined with cash to adjust for varying risk appetites) rather than an optimal portfolio based on historical averages. How can we anchor our predictive distribution to a portfolio that is well known to investors and already serves as a passive alternative, ie, the market portfolio?

Again, we use two independent sets of information to generate the predictive distribution.[26] Let us begin with the anchor: instead of assuming that returns come from a historical distribution $\boldsymbol{r} \sim$

$N(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Omega}})$, it is assumed that returns come from

$$r \sim N(\boldsymbol{\pi}, \tau\hat{\boldsymbol{\Omega}}) \tag{6.17}$$

where $\boldsymbol{\pi}$ denotes the equilibrium returns (those returns that would give the market portfolio if used in an optimisation) and τ reflects the level of belief in the equilibrium returns. The smaller τ is, the more weight is given to the equilibrium returns (historical uncertainty is scaled down). Investors' priors provide the second source of information about mean returns. It is assumed that the investor has some knowledge about the distribution of returns that allows them to forecast them according to

$$\boldsymbol{\mu}_{\text{prior}} = \boldsymbol{r} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Gamma}) \tag{6.18}$$

where $\boldsymbol{\Gamma}$ is the variance–covariance of forecast errors. Prior views are regarded as unbiased estimators of future returns, whereas forecasting errors are seen as pure noise. The elements of $\boldsymbol{\Gamma}$ can be easily determined if we already use a quantitative forecasting model. In that case, we can scale down the elements on the main diagonal by using the explained variance available from the respective forecasting models

$$\boldsymbol{\Gamma} = \begin{pmatrix} \sigma_1^2(1 - R_1^2) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_k^2(1 - R_k^2) \end{pmatrix} \tag{6.19}$$

In Equation 6.19 we implicitly assume that we forecast "true" alpha because only asset-specific risk (idiosyncratic risk) will show, by definition, zero correlation across assets. If we were to forecast total returns (or risk premiums), we would very likely find forecast errors to be positively correlated. However, keeping in mind that IC $= \sqrt{R^2}$, we can use the information coefficient (IC) – the correlation between a forecast return and the return actually realised – to link forecasting skills to the covariance matrix of forecasting errors. High skills (high IC) shrink the diagonal terms in $\boldsymbol{\Gamma}$ and thus give more weight to high skill forecasts.

Alternatively, we can use a more judgmental approach, transforming typical qualitative forecasts into a quantitative view. For example, an investor who is 90% sure that the return on the US market will be between 5% and 15% in fact believes the average return to be about 10% with a volatility of about 3.33% (assuming normality);

we can say this because we know that 90% of all data lie within three standard deviations $(10\% \pm 1.5 \times 3.33\%)$.

Analogous to Equations 6.11 and 6.14, the mean predictive return (refined forecast) can be calculated as a matrix-weighted average of two information sources[27]

$$\boldsymbol{\mu}_{\text{predictive}} = (\boldsymbol{\Gamma}^{-1} + \boldsymbol{\tau}^{-1}\hat{\boldsymbol{\Omega}}^{-1})^{-1}(\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}_{\text{prior}} + \boldsymbol{\tau}^{-1}\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{\pi}) \qquad (6.20)$$

where again $\boldsymbol{\tau}^{-1}\hat{\boldsymbol{\Omega}}^{-1}$ measures the precision (confidence) in the forecasts of equilibrium returns. The parameter $\boldsymbol{\tau}$ plays a similar role to T in Equation 6.14, and $\boldsymbol{\Gamma}^{-1}$ measures the precision of our prior views. Although a slightly different interpretation is given to the matrixes involved, Equations 6.14 and 6.20 have the same solution.

6.4.2 A three-asset example

A simple three-asset example can be used to provide further insight into the mechanics. Suppose we have three highly correlated markets, each with a correlation of 0.9 to the others, an allocation of one third and a volatility of 20%. We would expect a portfolio like this to generate 4% excess returns (hence the implied returns must be 4% each as the three assets do not differ in terms of risk characteristics). We are positive about asset one $(+5\%)$ and negative about assets two and three (both -1%). We also assume that our forecast errors are independent (though this is not always realistic if forecasts are derived from a common macroeconomic scenario). Assuming our forecasts are 6% for each asset and that forecast errors are the same size as asset volatilities, we get Equation 6.21 on page 182.

Although the range of forecasts is high, this is not reflected in the predictive means. Forecasts are automatically compressed to reflect the high correlation between assets. Hence, return estimates come more into line with the covariance matrix of asset returns, which will allow more realistic asset allocations.

Suppose instead that we have little confidence in our forecast, ie, that we make the diagonal terms in the matrix of forecast errors large, as shown in Equation 6.22 on page 182.

In this case we find the refined forecasts to be very close to the equilibrium forecast, with effectively all their variation removed.

6.4.3 Effect of varying $\boldsymbol{\tau}$

Finally, we can look at the effect of variations in τ, the parameter that reflects the level of belief in the equilibrium returns. Suppose

$$\left(\frac{\begin{pmatrix}400 & 0 & 0\\ 0 & 400 & 0\\ 0 & 0 & 400\end{pmatrix}^{-1}}{\begin{pmatrix}400 & 0 & 0\\ 0 & 400 & 0\\ 0 & 0 & 400\end{pmatrix}^{-1} + \tau^{-1}\begin{pmatrix}400 & 360 & 360\\ 360 & 400 & 360\\ 360 & 360 & 400\end{pmatrix}^{-1}}\right)\begin{pmatrix}5\\ -1\\ -1\end{pmatrix}$$

$$+\left(\frac{\tau^{-1}\begin{pmatrix}400 & 360 & 360\\ 360 & 400 & 360\\ 360 & 360 & 400\end{pmatrix}^{-1}}{\begin{pmatrix}400 & 0 & 0\\ 0 & 400 & 0\\ 0 & 0 & 400\end{pmatrix}^{-1} + \tau^{-1}\begin{pmatrix}400 & 360 & 360\\ 360 & 400 & 360\\ 360 & 360 & 400\end{pmatrix}^{-1}}\right)\begin{pmatrix}4\\ 4\\ 4\end{pmatrix}$$

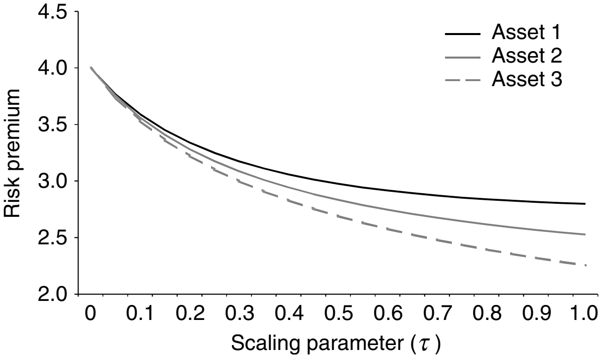
$$=\begin{pmatrix}2.15\\ 1.61\\ 1.61\end{pmatrix} \qquad (6.21)$$

$$\left(\frac{\begin{pmatrix}10{,}000 & 0 & 0\\ 0 & 10{,}000 & 0\\ 0 & 0 & 10{,}000\end{pmatrix}^{-1}}{\begin{pmatrix}10{,}000 & 0 & 0\\ 0 & 10{,}000 & 0\\ 0 & 0 & 10{,}000\end{pmatrix}^{-1} + \tau^{-1}\begin{pmatrix}400 & 360 & 360\\ 360 & 400 & 360\\ 360 & 360 & 400\end{pmatrix}^{-1}}\right)\begin{pmatrix}5\\ -1\\ -1\end{pmatrix}$$

$$+\left(\frac{\tau^{-1}\begin{pmatrix}400 & 360 & 360\\ 360 & 400 & 360\\ 360 & 360 & 400\end{pmatrix}^{-1}}{\begin{pmatrix}10{,}000 & 0 & 0\\ 0 & 10{,}000 & 0\\ 0 & 0 & 10{,}000\end{pmatrix}^{-1} + \tau^{-1}\begin{pmatrix}400 & 360 & 360\\ 360 & 400 & 360\\ 360 & 360 & 400\end{pmatrix}^{-1}}\right)\begin{pmatrix}4\\ 4\\ 4\end{pmatrix}$$

$$=\begin{pmatrix}3.71\\ 3.69\\ 3.69\end{pmatrix} \qquad (6.22)$$

Figure 6.7 Forecasts as a function of confidence in equilibrium returns



we forecast returns of 5% for asset one, 2% for asset two and −1% for asset three. Forecast errors are assumed to be uncorrelated and equal in size to asset volatility (20%).

Remember that if τ becomes smaller, this is equivalent to saying that our confidence in the equilibrium returns has increased as a reduced τ will scale down the covariance matrix of historical returns. In addition, as $\tau^{-1}\hat{\boldsymbol{\Omega}}^{-1}$ becomes larger, the weight attached to the equilibrium returns and $(\boldsymbol{\Gamma}^{-1} + \tau^{-1}\hat{\boldsymbol{\Omega}}^{-1})^{-1}(\tau^{-1}\hat{\boldsymbol{\Omega}}^{-1})$ also increases, approaching one. This effect is captured in Figure 6.7 for the 5%, 2% and −1% forecasts. As τ approaches zero, we arrive at the equilibrium forecasts of 4%. However, if our uncertainty about the equilibrium returns increases (ie, as τ approaches 1), the spread between the three forecasts will rise once again. Reduced uncertainty about the historical data will not lead back to the sample data but back to the equilibrium returns. This is a crucial difference from the general multivariate case considered in Section 6.3.

The Black–Litterman model has proved popular with investors as it allowed them for the first time to separate forecast (“where will the market go?”) from conviction (“how confident am I that my forecast is true?”).[28] Additionally, it anchors the solution into the well-known market portfolio and creates consistency between raw forecasts and sample covariance matrixes.

However, the B–L method will not use all the information available in historical means. It could well be argued that the model presented in this section puts an implicit prior (zero alpha with no uncertainty) on the capital asset pricing model (CAPM) as the true equilibrium model,[29] though this is not necessarily everybody's conviction.[30] Moreover, the model does not directly address parameter uncertainty, making Equation 6.14 a more complete model of reality as it incorporates the B–L model as a special case. Also, we have not yet dealt with problems arising from messy data. For example, how do we optimally combine time series with different starting and ending periods? The next section addresses this problem.

6.5 UNCERTAINTY ABOUT THE CAPM AS EQUILIBRIUM MODEL

Using implied returns implicitly assumes that there is no uncertainty about the equilibrium model and all assets are fairly priced, ie, mean–variance spanned. We can see this directly from the following argument. Suppose the CAPM is the true equilibrium model. Excess returns for all given assets are then calculated as the product of estimated market beta and the markets excess return

$$\boldsymbol{\mu} = \hat{\boldsymbol{\beta}}\hat{\mu}_{\mathrm{m}} \tag{6.23}$$

In a CAPM world, the (world-) market portfolio is the only benchmark asset and we can write the covariance matrix of the k-asset universe with the one and only benchmark asset as

$$\begin{bmatrix} \boldsymbol{\Omega} & \boldsymbol{\Omega w} \\ \boldsymbol{w}'\boldsymbol{\Omega} & \boldsymbol{w}'\boldsymbol{\Omega w} \end{bmatrix} \tag{6.24}$$

where $\boldsymbol{w}$ denotes the $k \times 1$ weight vector of individual assets included in the benchmark asset and $\boldsymbol{\beta} = \boldsymbol{\Omega w}/\boldsymbol{w}'\boldsymbol{\Omega w}$ describes the $k \times 1$ vector of asset betas relative to the benchmark asset.[31] The optimal portfolio is as usual given by

$$\begin{aligned} w^* &= \frac{1}{\lambda}\begin{bmatrix} \boldsymbol{\Omega} & \boldsymbol{\Omega w} \\ \boldsymbol{w}'\boldsymbol{\Omega} & \boldsymbol{w}'\boldsymbol{\Omega w} \end{bmatrix}^{-1}\begin{bmatrix} \boldsymbol{\beta}\mu_{\mathrm{m}} \\ \mu_{\mathrm{m}} \end{bmatrix} \\ &= \frac{1}{\lambda}\begin{bmatrix} \mathbf{0} \\ (\boldsymbol{w}'\boldsymbol{\Omega w})^{-1}\mu_{\mathrm{m}} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} \\ \mu_{\mathrm{m}}/\lambda\sigma_{\mathrm{m}}^2 \end{bmatrix} \end{aligned} \tag{6.25}$$

In other words, given zero alpha return expectations given in Equation 6.23 you would not add any of the assets to the benchmark portfolio. It already is the maximum Sharpe ratio portfolio for these return expectations. This is a rather dogmatic view. Suppose we allow uncertainty about "alpha", ie, deviations from Equation 6.23 and introduce a prior on alpha

$$p(\boldsymbol{\alpha} \mid \boldsymbol{\Omega}) = N(\mathbf{0}, \theta[\boldsymbol{\Omega} - \beta\sigma_{\mathrm{m}}^2\beta']) \tag{6.26}$$

The prior exhibits zero mean and its variance is shrunk via θ relative to the variability of the residuals of k linear CAPM regressions captured by $\boldsymbol{\Omega} - \beta\sigma_{\mathrm{m}}^2\beta'$. Here the main diagonal contains the variance of each regression residual, while the off-diagonal elements contain their covariance. After some tedious manipulations we get

$$\boldsymbol{\mu}_{\text{posterior}} = \hat{\boldsymbol{\beta}}\mu_{\mathrm{m}} + \hat{\boldsymbol{\alpha}}(1 - \omega) \tag{6.27}$$

where $\hat{\boldsymbol{\alpha}}$ equals $\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\beta}}\hat{\mu}_{\mathrm{m}}$ (ie, the constant in a linear regression of asset returns against the pricing factor) and shrinkage is defined as

$$\omega = \frac{1}{1 + T\theta/(\hat{\mu_{\mathrm{m}}}/\sigma_{\mathrm{m}})^2} \tag{6.28}$$

In the case where we have full belief in the asset pricing model, we set $\theta = 0$, which implies that $\omega = 1$ and in turn $\boldsymbol{\mu}_{\text{posterior}} = \hat{\boldsymbol{\beta}}\mu_{\mathrm{m}}$. This would lead us back to the benchmark portfolio. If, however, we have no belief in the asset pricing model, we set θ to become infinity. As a consequence we find $\omega = 0$ and $\boldsymbol{\mu}_{\text{posterior}} = \hat{\boldsymbol{\beta}}\mu_{\mathrm{m}} + \hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\mu}}$. The result would be a massive departure from the benchmark portfolio. Varying degrees of confidence in the CAPM as equilibrium model will lead to varying departures from the benchmark (market) portfolio.

A richer and computationally very efficient way to allow for uncertainty in the equilibrium model (as well as all asset betas) is to compute a Bayesian linear regression model of the form $r_i = \alpha_i + \beta_i r_{\mathrm{m}} + \varepsilon_i$ with priors on "alpha" and "beta". This could be easily extended to a multivariate Bayesian regression (to accommodate for the Fama–French factors) and is hard coded in many software packages (S$^+$Bayes, for example).[32]

6.6 INTRODUCING FACTOR AND ALPHA BETS IN THE B–L MODEL

At this point we best write the B–L model in its more general form, where the updated posterior means are given by

$$\boldsymbol{\mu}_{\text{predictive}} = ((\tau\boldsymbol{\Omega})^{-1} + \boldsymbol{P}'\boldsymbol{\Gamma}^{-1}\boldsymbol{P})^{-1}((\tau\boldsymbol{\Omega})^{-1}\boldsymbol{\pi} + \boldsymbol{P}'\boldsymbol{\Gamma}^{-1}\boldsymbol{q})^{-1} \tag{6.29}$$

Table 6.2 Traditional betting structure in B–L model

Intended bet	"Pick matrix" representation
Directional bets on assets one and two	$\boldsymbol{P} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \end{bmatrix}$
Spread bets on asset one versus asset two and asset two versus asset three	$\boldsymbol{P} = \begin{bmatrix} & & & \\ 1 & -1 & \cdots & 0 \\ \cdots & 1 & -1 & \cdots \end{bmatrix}$
Basket bets (asset one versus equal weighted basket of remaining assets)	$\boldsymbol{P} = \begin{bmatrix} 1 & -1/(k-1) & \cdots & -1/(k-1) \end{bmatrix}$

Table 6.3 Number of views in the augmented B–L model

Type of view	Number of bets (rows)	Number of assets (columns)	Pick matrix
Total return	n_k	k	$\boldsymbol{P}$
Factor	n_f	f	$\boldsymbol{P}_f$
Alpha	n_α	k	$\boldsymbol{P}_\alpha$

Here we introduce a "pick matrix", $\boldsymbol{P}$, which specifies our bet structure. It is of dimension $n_k \times k$, where n_k denotes the number of bets we undertake on the total return of k assets. Various betting structures can be supported (Table 6.2).

Note that in the previous section we implicitly set $\boldsymbol{P}$ to become a diagonal $k \times k$ matrix with 1s on the main diagonal. All other notation remains the same with the exception that we exchanged $\boldsymbol{\mu}_{\text{prior}}$ with the more conventional $\boldsymbol{q}$ (to represent our views).

Suppose instead we live in a multifactor world, ie, returns on a given asset are also driven by style, country or industry factors. Let us generically assume an f factor world. How can we use to use the BL model to forecast factor returns and alpha returns as in its current form we can only use it for total returns? Suppose we want to implement the bets in Table 6.3.

Cheung (2009) has shown that we can maintain the structure of Equation 6.29 and merely augment the matrixes involved for the new "alpha" and "factor beta" bets

$$\boldsymbol{\mu}_{\mathrm{a,predictive}} = ((\tau\boldsymbol{\Omega}_{\mathrm{a}})^{-1} + \boldsymbol{P}'_{\mathrm{a}}\boldsymbol{\Gamma}_{\mathrm{a}}^{-1}\boldsymbol{P}_{\mathrm{a}})^{-1}((\tau\boldsymbol{\Omega}_{\mathrm{a}})^{-1}\boldsymbol{\pi}_{\mathrm{a}} + \boldsymbol{P}'_{\mathrm{a}}\boldsymbol{\Gamma}_{\mathrm{a}}^{-1}\boldsymbol{q}_{\mathrm{a}})^{-1} \tag{6.30}$$

The matrix augmentation associated with Equation 6.30 is straightforward. Given we have three sets of views we also have three sets of forecasting errors, ie

$$\boldsymbol{\Gamma}_{\mathrm{a}} = \begin{bmatrix} \boldsymbol{\Gamma} & 0 & 0 \\ 0 & \boldsymbol{\Gamma}_F & 0 \\ 0 & 0 & \boldsymbol{\Gamma}_\alpha \end{bmatrix}_{(n_k+n_f+n_\alpha)\times(n_k+n_f+n_\alpha)} \tag{6.31}$$

Our matrix of asset risks now needs to contain both factors, as well as asset-specific risks

$$\boldsymbol{\Omega}_{\mathrm{a}} = \begin{bmatrix} \boldsymbol{\Omega} & \boldsymbol{B}\boldsymbol{\Omega}_{FF} \\ \boldsymbol{\Omega}_{FF}\boldsymbol{B}' & \boldsymbol{\Omega}_{FF} \end{bmatrix} = \begin{bmatrix} \boldsymbol{B}\boldsymbol{\Omega}_{FF}\boldsymbol{B}' + \boldsymbol{\Omega}_{\alpha\alpha} & \boldsymbol{B}\boldsymbol{\Omega}_{FF} \\ \boldsymbol{\Omega}_{FF}\boldsymbol{B}' & \boldsymbol{\Omega}_{FF} \end{bmatrix} \tag{6.32}$$

Here $\boldsymbol{B}$ represents a $k \times f$ matrix of factor exposures. For example, the first row represents the betas for the first asset against all factors. In a four-factor world (value/HML, size/SMB, market/MKT, momentum/MOM), this would, for example, amount to

$$\boldsymbol{B} = \begin{bmatrix} \beta_{1,\mathrm{MKT}} & \beta_{1,\mathrm{HML}} & \beta_{1,\mathrm{SMB}} & \beta_{1,\mathrm{MOM}} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{k,\mathrm{MKT}} & \beta_{K,\mathrm{HML}} & \beta_{K,\mathrm{SMB}} & \beta_{K,\mathrm{MOM}} \end{bmatrix} \tag{6.33}$$

The implied equilibrium returns are given from

$$\boldsymbol{\pi}_{\mathrm{a}} = \begin{bmatrix} \boldsymbol{\pi} \\ \boldsymbol{\pi}_f \end{bmatrix} = (\mu - r)\begin{bmatrix} (\boldsymbol{B}\boldsymbol{\Omega}_{FF}\boldsymbol{B}' + \boldsymbol{\Omega}_{\alpha\alpha})\boldsymbol{w}_{\mathrm{m}}/\sigma_{\mathrm{m}}^2 \\ \boldsymbol{\Omega}_{FF}\boldsymbol{B}'\boldsymbol{w}_{\mathrm{m}}/\sigma_{\mathrm{m}}^2 \end{bmatrix} \tag{6.34}$$

This is the usual beta interpretation, ie, implied returns are given by beta times market risk premium, where $(\boldsymbol{B}\boldsymbol{\Omega}_{FF}\boldsymbol{B}' + \boldsymbol{\Omega}_{\alpha\alpha})\boldsymbol{w}_{\mathrm{m}}/\sigma_{\mathrm{m}}^2$ denotes a $k \times 1$ vector of asset betas versus the market portfolio and $\boldsymbol{\Omega}_{FF}\boldsymbol{B}'\boldsymbol{w}_{\mathrm{m}}/\sigma_{\mathrm{m}}^2$ represents an $f \times 1$ vector of factor betas with the same market portfolio (represented by the $k \times 1$ vector $\boldsymbol{w}_{\mathrm{m}}$). In the special case of the CAPM as a one-factor model we see

$$\frac{\boldsymbol{\Omega}_{FF}\boldsymbol{B}'\boldsymbol{w}_{\mathrm{m}}}{\sigma_{\mathrm{m}}^2} = \frac{\sigma_{\mathrm{m}}^2}{\sigma_{\mathrm{m}}^2}\sum_{i=1}^{k} w_i\beta_i = 1 \tag{6.35}$$

which is what we would expect. We can easily check the implied asset and factor returns. If returns obey Equation 6.35, all information is already in $\boldsymbol{\pi}$ while $\boldsymbol{\pi}_f$ is perfectly spanned

$$\begin{aligned}\begin{bmatrix} \boldsymbol{w}_{\text{m}} \\ \boldsymbol{w}_f \end{bmatrix} &= \lambda^{-1}\boldsymbol{\Omega}_{\text{a}}^{-1}\begin{bmatrix} \boldsymbol{\pi} \\ \boldsymbol{\pi}_{\text{a}} \end{bmatrix} \\ &= \lambda^{-1}\begin{bmatrix} \boldsymbol{\Omega} & \boldsymbol{B}\boldsymbol{\Omega}_{FF} \\ \boldsymbol{\Omega}_{FF}\boldsymbol{B}' & \boldsymbol{\Omega}_{FF} \end{bmatrix}^{-1}\begin{bmatrix} \boldsymbol{\Omega}\boldsymbol{w}_{\text{m}}/\sigma_{\text{m}}^2 \\ \boldsymbol{\Omega}_{FF}\boldsymbol{B}'\boldsymbol{w}_{\text{m}}/\sigma_{\text{m}}^2 \end{bmatrix}(\mu - c) \end{aligned} \quad (6.36)$$

In other words the "factor assets" are redundant and should carry zero weight, ie, $\boldsymbol{w}_f = \boldsymbol{0}$. We will use the partitioned inverse of $\boldsymbol{\Omega}_{\text{a}}$ to get

$$\begin{aligned}\boldsymbol{w}_f = \Big[-\boldsymbol{\Omega}_{FF}^{-1}\boldsymbol{\Omega}_{FF}\boldsymbol{B}'\Delta^{-1} \quad & \boldsymbol{\Omega}_{FF}^{-1} + \boldsymbol{\Omega}_{FF}^{-1}\boldsymbol{\Omega}_{FF}\boldsymbol{B}'\Delta^{-1}\boldsymbol{B}\boldsymbol{\Omega}_{FF}\boldsymbol{\Omega}_{FF}^{-1}\Big] \\ & \times \begin{bmatrix} \boldsymbol{\Omega}\boldsymbol{w}_{\text{m}}/\sigma_{\text{m}}^2 \\ \boldsymbol{\Omega}_{FF}\boldsymbol{B}'\boldsymbol{w}_{\text{m}}/\sigma_{\text{m}}^2 \end{bmatrix}(\mu - c)\lambda^{-1} \end{aligned} \quad (6.37)$$

where $\Delta = (\boldsymbol{\Omega} - \boldsymbol{B}\boldsymbol{\Omega}_{FF}\boldsymbol{\Omega}_{FF}^{-1}\boldsymbol{\Omega}_{FF}\boldsymbol{B}')^{-1} = \boldsymbol{\Omega}_{\alpha\alpha}^{-1}$. After expanding Equation 6.37 we arrive at

$$\begin{aligned}\boldsymbol{w}_f &= \frac{-\boldsymbol{B}'\Delta^{-1}\boldsymbol{B}\boldsymbol{\Omega}_{FF}\boldsymbol{B}'\boldsymbol{w}_{\text{m}} - \boldsymbol{B}'\Delta^{-1}\boldsymbol{\Omega}_{\alpha\alpha}\boldsymbol{w}_{\text{m}}}{\sigma_{\text{m}}^2}(\mu - c) \\ &\quad + \frac{\boldsymbol{B}'\boldsymbol{w}_{\text{m}} + \boldsymbol{B}'\Delta^{-1}\boldsymbol{B}\boldsymbol{\Omega}_{FF}\boldsymbol{B}'\boldsymbol{w}_{\text{m}}}{\sigma_{\text{m}}^2}(\mu - c) \\ &= \frac{-\boldsymbol{B}'\Delta^{-1}\boldsymbol{\Omega}_{\alpha\alpha}\boldsymbol{w}_{\text{m}} + \boldsymbol{B}'\boldsymbol{w}_{\text{m}}}{\sigma_{\text{m}}^2}(\mu - c) \\ &= \frac{-\boldsymbol{B}'\boldsymbol{\Omega}_{\alpha\alpha}^{-1}\boldsymbol{\Omega}_{\alpha\alpha}\boldsymbol{w}_{\text{m}} + \boldsymbol{B}'\boldsymbol{w}_{\text{m}}}{\sigma_{\text{m}}^2}(\mu - c) \\ &= 0 \end{aligned} \quad (6.38)$$

The implied returns (Equation 6.34) lead us back to the market portfolio as they should. Next we look closer at our augmented pick matrix that describes our new betting structure. It becomes

$$\boldsymbol{P}_{\text{a}} = \begin{bmatrix} \boldsymbol{P} & 0 \\ 0 & \boldsymbol{P}_f \\ \boldsymbol{P}_\alpha & -\boldsymbol{P}_\alpha\boldsymbol{B} \end{bmatrix} \quad (6.39)$$

The matrix components are described in Table 6.3. With the ingredients above we can now implement alpha as well as beta (factor) bets in the B–L framework via forecasts $\boldsymbol{q}_{\text{a}}$. The usual B–L logic applies. A positive forecast on a single factor (view return larger than equilibrium return) will make this factor more attractive but will also make all stocks with exposure to this factor more attractive. So we can trade our factor forecast even if we could not trade the factor portfolios directly. While this logic is intuitive the framework by Cheung (2009) looks a bit daunting. For most practical purposes the author would recommend a simpler framework: just introduce factors you want to trade on as additional assets in the traditional BL approach. In other words create a characteristic portfolio and include the time series as additional assets in Equation 6.29. This is what practitioners did before the Cheung paper and it is literally equivalent. Also note that we do not necessarily need the above mathematics. Just use regressions coefficients instead of $(\boldsymbol{B\Omega}_{FF}\boldsymbol{B}' + \boldsymbol{\Omega}_{\alpha\alpha})\boldsymbol{w}_{\mathrm{m}}/\sigma_{\mathrm{m}}^2$ (regression of market portfolio return on factor return) and $\boldsymbol{\Omega}_{FF}\boldsymbol{B}'\boldsymbol{w}_{\mathrm{m}}/\sigma_{\mathrm{m}}^2$ (regression of market portfolio return on factor return).

6.7 THE USE OF HIERARCHICAL PRIORS IN MANAGER ALLOCATION

Suppose you are given a set of investments (for example, the benchmark excess returns of a set of hedge fund managers) for a given sample period. For each asset manager you measure a historical sample alpha $\bar{\alpha}_i$ with estimation error $\sigma_{\alpha_i}^2$ (manager variance divided by the number of available observations) for $i = 1, \dots, k$ investments. Suppose furthermore that you have few *a priori* views (or little insight what the managers actually do) why they should have different expected returns, yet you do not want to ignore the information in the data. This describes a familiar situation for analysts being asked to arrive at an optimal (alpha) allocation for a set of cash benchmarked active funds. Note that the following Bayesian procedure remains technically correct even if your subjective priors are totally wrong, ie, even if you ignore valuable information with regards to the managers strategies. Our hierarchical model looks like

$$\alpha_i \sim N(\alpha, \tau^2) \tag{6.40}$$

$$\bar{\alpha}_i \sim N(\alpha_i, \sigma_{\alpha_i}^2) \tag{6.41}$$

Equation 6.40 states that the individual alphas are random draws from the distribution for an unknown "group" alpha (we could, for example, view it as the inherent skill in all hedge fund managers of a given style, ie, their grand mean). The dispersion around this unknown group alpha is given by τ^2 (also unknown). In accordance with our ignorance, we set the prior distribution on τ to be uniform, ie, $p(\tau) \propto 1$. Each sample alpha in turn differs from the true alpha by its estimation error.

The estimation of the model follows a three-step approach; see Gelman *et al* (2003, p. 138) for technical details. Here we will focus on the implementation.

Step 1. Numerically evaluate the posterior distribution for τ given the observed dispersion in the sample data. It was shown earlier that

$$p(\tau \mid \text{data}) \propto p(\tau) V_{\mu}^{1/2} \prod_{i=1}^{k} (\sigma_{\bar{\alpha}_i}^2 + \tau^2)^{-1/2} \exp\left(-\frac{1}{2} \frac{(\bar{\alpha}_i - \bar{\alpha})^2}{\sigma_{\bar{\alpha}_i}^2 + \tau^2} \right) \tag{6.42}$$

$$\bar{\alpha} = \frac{\sum_{i=1}^{k} \bar{\alpha}_i / (\sigma_{\bar{\alpha}_i}^2 + \tau^2)}{\sum_{i=1}^{k} 1/(\sigma_{\bar{\alpha}_i}^2 + \tau^2)} \tag{6.43}$$

$$V_{\alpha}^{-1} = \sum_{i=1}^{k} \frac{1}{\sigma_{\bar{\alpha}_i}^2 + \tau^2} \tag{6.44}$$

Note the proportionality sign in Equation 6.42 and the fact that, given Equations 6.43 and 6.44, Equation 6.42 is a function of the data and τ alone. As the data is given we can evaluate Equation 6.42 on a grid, ie, calculate Equation 6.42 for different values of τ. Finally, we rescale all values of Equation 6.42 such that they form a proper probability distribution and draw $n = 1, \dots, N$ realisations for τ.

Step 2. For each sampled τ we can now draw the "corresponding" level of overall α according to

$$\alpha \sim N(\bar{\alpha}, V_{\alpha}^{-1}) \tag{6.45}$$

where $\bar{\alpha}$, V_{α}^{-1} are given in Equations 6.43 and 6.44, ie, they are functions of τ. This results in N realisations for α.

Step 3. For each sampled α and τ we now draw another N realisations for the unknown α_i, by drawing individually (ie, for each $i = 1, \dots, k$) from

$$\alpha_i \sim N\left(\left(\frac{\bar{\alpha}_i}{\sigma_{\bar{\alpha}_i}^2} + \frac{\alpha}{\tau^2} \right) \left(\frac{1}{\sigma_{\bar{\alpha}_i}^2} + \frac{1}{\tau^2} \right)^{-1}, \left(\frac{1}{\sigma_{\bar{\alpha}_i}^2} + \frac{1}{\tau^2} \right)^{-1} \right) \tag{6.46}$$

Table 6.4 Sample data for hierarchical prior example

	Manager 1	Manager 2	Manager 3	Manager 4
α_i (%)	20	10	10	−10
σ_{α_i} (%)	10	10	10	10

Clearly, this will shrink mean alphas towards α dependent on the relative precision of the "grand mean" versus the individual sample means.

After step 3 we arrive at a large set of sampled data, ie, $N \times (k+2)$ data points, for which we can now calculate marginal distributions as well as mean estimates

$$\begin{aligned} &\tau_1^2, \alpha_1, \alpha_{11}, \alpha_{12}, \dots, \alpha_{1k} \\ &\tau_2^2, \alpha_2, \alpha_{21}, \alpha_{22}, \dots, \alpha_{2k} \\ &\quad\vdots \\ &\tau_n^2, \alpha_n, \alpha_{n1}, \alpha_{n2}, \dots, \alpha_{nk} \\ &\quad\vdots \\ &\tau_N^2, \alpha_N, \alpha_{N1}, \alpha_{N2}, \dots, \alpha_{Nk} \end{aligned} \tag{6.47}$$

The posterior mean for the alpha of the nth manager can, for example, be found from $\alpha_i^h = -(1/N)\sum_{n=1}^{N} \alpha_{ni}$ given the data above. This procedure is your best shot if you are truly ignorant, ie, you view all investments as exchangeable. To put it differently, a situation where we have no idea what factors drive their relative performance.

In order to get a better understanding and offer the reader an opportunity to check their own code against our results, we suggest the following situation as outlined in Table 6.4. It contains "alphas" and their estimation errors. There is no requirement for all time series have the same number of data points or for the estimation error to be a statistical estimate.

First we model $p(\tau \mid \text{data})$ as given by Equation 6.42. The results are pictured in Figure 6.8. The modus of the distributions for τ is about 0.1 with an expected value of about 0.15. What determines its shape and dispersion? If the estimation error for individual manager alphas is very small all variation, all variation must come from Equation 6.40, ie, the distribution for τ will peak at a relatively large value, while it will find its peak close to zero if the estimation error

Figure 6.8 Scaled posterior density for τ according to Equation 6.42

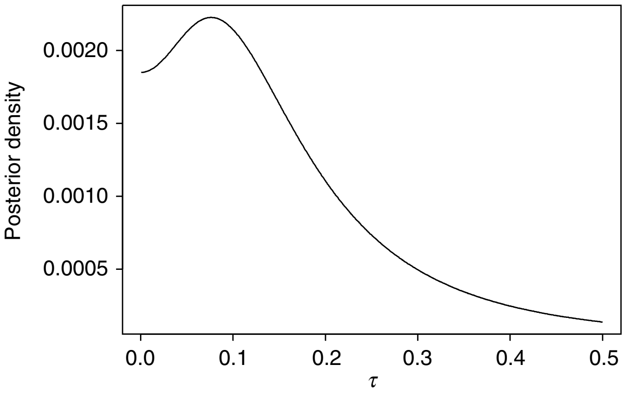


Table 6.5 Average alphas after shrinkage

	Manager 1	Manager 2	Manager 3	Manager 4
α_i^h (%)	13.8	8.9	8.9	−0.01

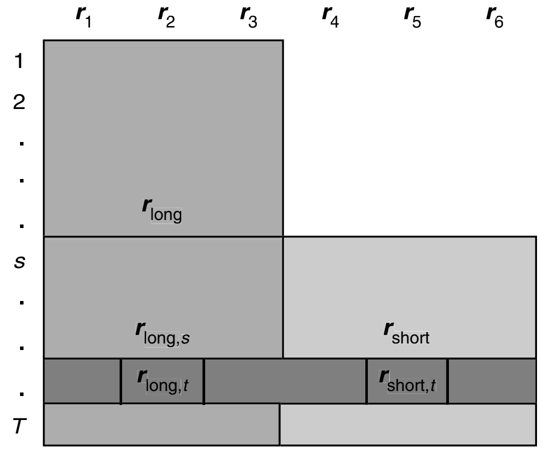
is large, ie, the variation is more likely to come from Equation 6.41. While we could use the mean estimate for τ directly and continue with Equations 6.45 and 6.46, this would ignore the considerable parameter uncertainty in our estimates. Instead we sample all data in Equation 6.47 for $N = 1{,}000$, ie, we first sample Equation 6.45 and then Equation 6.46 for $i = 1, \dots, 4$. Average alphas after shrinkage are shown in Table 6.5.

Note that the alpha estimate with the highest estimation error is shrunk most. This is what we would expect. Given that the expected value for τ was considerably large (0.15) with some uncertainty around it, the shrinkage is relatively modest.

6.8 TIME SERIES OF DIFFERENT LENGTHS

Asset allocation and portfolio construction exercises are based on the use of historical data. Typically, however, the availability of this data differs across asset classes:[33] return series for developed equity markets reach back to the early 1970s (MSCI data); data on government bonds (Salomon Brothers data) and emerging market equities (IFC data) date back to the mid-1980s; data on emerging market debt

Figure 6.9 Data structure for time series of different lengths



reach back to the early 1990s (JPM data); and data on corporate debt only go back to the mid-1990s (Merrill Lynch data).[34] Asset allocators attempting to deal with combinations of these asset classes face the problem of how to handle return series of differing lengths. Should they estimate means and volatilities separately (using the maximum number of observations for each series) and calculate correlations on the overlapping data set? Or should they truncate the longer series to the length of the shortest so that a rectangular data matrix can be generated? How can we apply Bayesian analysis dealing with estimation risk to return series of different lengths? This section looks at these issues in detail.[35]

Suppose we have a non-rectangular data matrix of risk premiums (one with time series of differing lengths) like that in Figure 6.9. We see three long series (T observations reaching from time 1 to T) and three short series (S observations, where $S = T - s + 1$ observations reaching from s to T). Long and short series are combined in $T \times 3$ and $S \times 3$ data matrixes called $\boldsymbol{r}_{\text{long}}$ and $\boldsymbol{r}_{\text{short}}$. It is assumed that the overlapping S data points come from a joint multivariate normal distribution (indicated by S). Hence, $\boldsymbol{r}_{\text{long},S}$ ($S \times 3$ data matrix) denotes the overlapping part of $\boldsymbol{r}_{\text{long}}$. An estimator that uses only the overlapping data (the rectangular data set that stretches from s to T and $\boldsymbol{r}_1$ to $\boldsymbol{r}_6$) is called a truncated estimator. These are just the normal sample estimators applied to the overlapping data set, and they can be written as follows

$$\left.\begin{aligned}
\hat{\boldsymbol{\mu}}^{\text{truncated}}_{\text{long}} &= \frac{1}{S}\boldsymbol{r}'_{\text{long},S}\mathbf{1}_S \\
\hat{\boldsymbol{\mu}}^{\text{truncated}}_{\text{short}} &= \frac{1}{S}\boldsymbol{r}'_{\text{short}}\mathbf{1}_S \\
\hat{\boldsymbol{\Omega}}^{\text{truncated}}_{\text{long, long}} &= \frac{1}{S}(\boldsymbol{r}_{\text{long},S} - \mathbf{1}_S\hat{\boldsymbol{\mu}}'^{\text{truncated}}_{\text{short}})'(\boldsymbol{r}_{\text{long},S} - \mathbf{1}_S\hat{\boldsymbol{\mu}}'^{\text{truncated}}_{\text{short}}) \\
&\vdots \\
\hat{\boldsymbol{\Omega}}^{\text{truncated}}_{\text{short, short}} &= \frac{1}{S}(\boldsymbol{r}_{\text{short},S} - \mathbf{1}_S\hat{\boldsymbol{\mu}}'^{\text{truncated}}_{\text{short}})'(\boldsymbol{r}_{\text{short},S} - \mathbf{1}_S\hat{\boldsymbol{\mu}}'^{\text{truncated}}_{\text{short}})
\end{aligned}\right\} \tag{6.48}$$

where $\mathbf{1}_S$ denotes an $S \times 1$ vector of 1s. The truncated-sample estimator ignores the information available in the non-overlapping part of the longer time series.

6.8.1 Combined-sample estimate

How would a combined-sample maximum likelihood (ML) estimator use the available information? First we have to recognise the dependencies between the long and the short series. Zero correlation between these series would imply that there is no information in the long series that could be used to estimate moments for the short series. Each of the short series is regressed against all of the three long series to capture dependencies

$$\left.\begin{aligned}
r_{4,t} &= \alpha_4 + \beta_{41}r_{1,t} + \beta_{42}r_{2,t} + \beta_{43}r_{3,t} + \varepsilon_{4,t} \quad \forall t = s,\dots,T \\
r_{5,t} &= \alpha_5 + \beta_{51}r_{1,t} + \beta_{52}r_{2,t} + \beta_{53}r_{3,t} + \varepsilon_{5,t} \quad \forall t = s,\dots,T \\
r_{6,t} &= \alpha_6 + \beta_{61}r_{1,t} + \beta_{62}r_{2,t} + \beta_{63}r_{3,t} + \varepsilon_{6,t} \quad \forall t = s,\dots,T
\end{aligned}\right\} \tag{6.49}$$

The regression coefficients are then summarised in a matrix we will call B, all the elements of which are βs

$$\boldsymbol{B}_{3\times3} = \begin{bmatrix} \hat{\beta}_{41} & \hat{\beta}_{51} & \hat{\beta}_{61} \\ \hat{\beta}_{42} & \hat{\beta}_{52} & \hat{\beta}_{62} \\ \hat{\beta}_{43} & \hat{\beta}_{53} & \hat{\beta}_{63} \end{bmatrix} \tag{6.50}$$

while the covariance of the residual terms (ie, the covariation between short series not explained by long series) is given as

$$\hat{\boldsymbol{\Omega}}_{\varepsilon\varepsilon} = \begin{bmatrix} \operatorname{cov}(\hat{\boldsymbol{\varepsilon}}_4, \hat{\boldsymbol{\varepsilon}}_4) & \operatorname{cov}(\hat{\boldsymbol{\varepsilon}}_4, \hat{\boldsymbol{\varepsilon}}_5) & \operatorname{cov}(\hat{\boldsymbol{\varepsilon}}_4, \hat{\boldsymbol{\varepsilon}}_6) \\ \operatorname{cov}(\hat{\boldsymbol{\varepsilon}}_5, \hat{\boldsymbol{\varepsilon}}_4) & \operatorname{cov}(\hat{\boldsymbol{\varepsilon}}_5, \hat{\boldsymbol{\varepsilon}}_5) & \operatorname{cov}(\hat{\boldsymbol{\varepsilon}}_5, \hat{\boldsymbol{\varepsilon}}_6) \\ \operatorname{cov}(\hat{\boldsymbol{\varepsilon}}_6, \hat{\boldsymbol{\varepsilon}}_4) & \operatorname{cov}(\hat{\boldsymbol{\varepsilon}}_6, \hat{\boldsymbol{\varepsilon}}_5) & \operatorname{cov}(\hat{\boldsymbol{\varepsilon}}_6, \hat{\boldsymbol{\varepsilon}}_6) \end{bmatrix} \tag{6.51}$$

This data can be used to decompose the total volatility of the short series into a part of total volatility explained by the long series (which can later be used to transform information from the long to the short series) and a part attributable only to the short series

$$\hat{\boldsymbol{\Omega}}^{\text{truncated}}_{\text{short, short}} = \underbrace{\hat{\boldsymbol{\Omega}}^{\text{truncated}}_{\varepsilon\varepsilon}}_{\substack{\text{Unexplained}\\ \text{variation}}} + \underbrace{\boldsymbol{B}\hat{\boldsymbol{\Omega}}^{\text{truncated}}_{\text{long, long}}\boldsymbol{B}}_{\substack{\text{Explained}\\ \text{variation}}} \tag{6.52}$$

The maximum likelihood estimates for mean and variance for the long series should cause little surprise, being just the normal sample estimates applied to the complete history as these series contain the most information anyway

$$\hat{\boldsymbol{\mu}}^{\text{ML}}_{\text{long}} = \frac{1}{T}\boldsymbol{r}'_{\text{long}}\mathbf{1}_S \tag{6.53}$$

$$\hat{\boldsymbol{\Omega}}^{\text{ML}}_{\text{long, long}} = \frac{1}{T}(\boldsymbol{r}_{\text{long}} - \mathbf{1}_T\hat{\boldsymbol{\mu}}'^{\text{ML}}_{\text{long}})'(\boldsymbol{r}_{\text{long}} - \mathbf{1}_T\hat{\boldsymbol{\mu}}'^{\text{ML}}_{\text{long}}) \tag{6.54}$$

As a direct effect of using more data the sample error decreases but, as already mentioned in the previous chapter, this will only be of any help if the data is drawn from a stationary process; it is the mean and variance estimators for the shorter series that change. The maximum likelihood estimator for the short series, $\hat{\boldsymbol{\mu}}^{\text{ML}}_{\text{short}}$, allocates the difference in the mean estimates of maximum likelihood and truncated estimator for the long series. This difference is due to sample error as the longer series estimate is assumed to be closer to the true mean, and it is distributed according to the respective βs

$$\hat{\boldsymbol{\mu}}^{\text{ML}}_{\text{short}} = \hat{\boldsymbol{\mu}}^{\text{truncated}}_{\text{short}} + \hat{\boldsymbol{B}}(\hat{\boldsymbol{\mu}}^{\text{ML}}_{\text{long}} - \hat{\boldsymbol{\mu}}^{\text{truncated}}_{\text{long}}) \tag{6.55}$$

If, for example, two series have a β of 0.5, only 50% of this difference is added to the truncated estimator (see the example below). The same logic applies to the updated variance estimate

$$\hat{\boldsymbol{\mu}}^{\text{ML}}_{\text{short, short}} = \hat{\boldsymbol{\Omega}}^{\text{truncated}}_{\varepsilon\varepsilon} + \boldsymbol{B}\hat{\boldsymbol{\Omega}}^{\text{ML}}_{\text{long, long}}\boldsymbol{B}' \tag{6.56}$$

The maximum likelihood adds the explained variation in short series movements to the unexplained variation but uses the whole data set to produce an estimate for the explained variation. Comparing Equations 6.52 and 6.56 shows this directly. This also extends to the covariance of long and short series

$$\hat{\boldsymbol{\Omega}}^{\text{ML}}_{\text{long, short}} = \boldsymbol{B}\hat{\boldsymbol{\Omega}}^{\text{ML}}_{\text{long, long}} \tag{6.57}$$

If volatility has been unusually low in the recent past (and short as well as long series show too little risk) but much higher over the longer historical record, this method will leverage up the volatility estimates according to the co-movements with the longer series. We can summarise by saying that the use of a broader set of historical information not only improves the estimation of moments of the longer series (less sampling error), but it also changes the mean estimates for the shorter series.[36]

6.8.2 Introducing estimation error

So far we have just used sample information, but how would our estimates change if we were to use a non-informative prior? As we know from Section 6.1, the maximum likelihood mean estimates will not change. However, in that instance we included the case of unequal time-series lengths because the maximum likelihood estimator differs from the truncated estimator (usually used in Markowitz applications), and this difference will change the optimal allocations. The Bayesian covariance matrix will simply be a leveraged version of the maximum likelihood estimates (compare with Section 6.8.1)

$$\left.\begin{aligned} \hat{\boldsymbol{\Omega}}^{\text{predictive}}_{\text{long, long}} &= \frac{T+1}{T-k-2}\hat{\boldsymbol{\Omega}}^{\text{ML}}_{\text{long, long}} \\ \hat{\boldsymbol{\Omega}}^{\text{predictive}}_{\text{long, short}} &= \frac{T+1}{T-k-2}\hat{\boldsymbol{\Omega}}^{\text{ML}}_{\text{long, short}} \end{aligned}\right\} \tag{6.58}$$

The more data is available for T, the smaller the difference between sample estimate (historical covariance) and predictive covariance. A slightly more complicated expression must be used to generate an expression for the covariance matrix of short series. However, in this expression, given below, both matrixes are leveraged up again

$$\begin{aligned} &\hat{\boldsymbol{\Omega}}^{\text{predictive}}_{\text{short, short}} \\ &\quad = \Delta\hat{\boldsymbol{\Omega}}^{\text{truncated}}_{\varepsilon\varepsilon} + \frac{T+1}{T-N-2}\boldsymbol{B}\hat{\boldsymbol{\Omega}}^{\text{ML}}_{\text{long, long}}\boldsymbol{B}' \end{aligned} \tag{6.59}$$

$$\begin{aligned} \Delta &= \left(\frac{S}{S-k_{\text{short}}-2}\right) \\ &\quad \times \left(1 + \frac{1}{S}\left[1 + \frac{T+1}{T-k-2}\,\text{trace}((\hat{\boldsymbol{\Omega}}^{\text{truncated}}_{\text{long, long}})^{-1}\hat{\boldsymbol{\Omega}}^{\text{ML}}_{\text{long, long}}) \right.\right. \\ &\qquad\qquad \left.\left. + (\hat{\boldsymbol{\mu}}^{\text{ML}}_{\text{long}} - \hat{\boldsymbol{\mu}}^{\text{truncated}}_{\text{long}})'\hat{\boldsymbol{\Omega}}^{-1}_{\text{long, long},S}(\hat{\boldsymbol{\mu}}^{\text{ML}}_{\text{long}} - \hat{\boldsymbol{\mu}}^{\text{truncated}}_{\text{long}})\right]\right) \end{aligned} \tag{6.60}$$

Table 6.6 Data for missing return series example

Quartal	European equities (%)	Latin American equities	Quartal	European equities (%)	Latin American equities (%)
Q3 93	8.21	—	Q2 97	8.44	20.7
Q4 93	9.01	—	Q3 97	7.81	4.2
Q1 94	−1.99	—	Q4 97	−0.36	−10.2
Q2 94	−1.86	—	Q1 98	19.84	0.2
Q3 94	3.71	—	Q2 98	4.75	−18.4
Q4 94	0.36	—	Q3 98	−14.77	−24.4
Q1 95	5.55	—	Q4 98	18.26	6.9
Q2 95	5.69	—	Q1 99	−2.49	7.5
Q3 95	3.62	—	Q2 99	0.70	16.9
Q4 95	2.84	—	Q3 99	0.76	−9.4
Q1 96	3.15	—	Q4 99	16.97	35.6
Q2 96	2.06	—	Q1 00	−0.26	8.7
Q3 96	3.24	—	Q2 00	−3.50	−6.4
Q4 96	9.02	—	Q3 00	−7.63	−6.7
Q1 97	4.37	—	Q4 00	1.61	−9.2

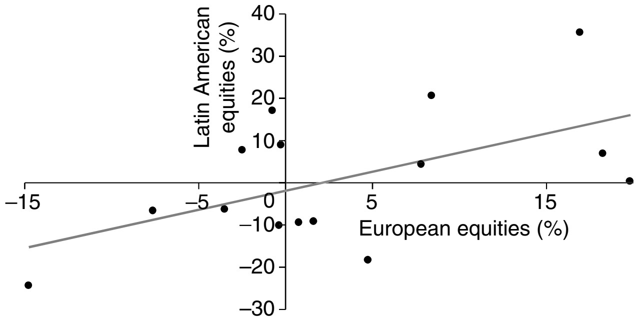
Table 6.7 Results for data in Table 6.6

	Truncated-sample estimator		Maximum likelihood		Bayesian estimates	
	European equities	Latin American equities	European equities	Latin American equities	European equities	Latin American equities
μ	12.99%	4.27%	14.10%	5.26%	14.10%	5.26%
σ	19.34%	31.44%	14.26%	29.14%	15.28%	35.43%
SR_i	0.67	0.14	0.99	0.18	0.92	0.15

Although the mathematics is straightforward, it should be stressed once more that we assume a constant marginal distribution of the long-history assets and constant correlation.

A simple two-asset example will illustrate the core principles. Return data for European and Latin American equities is set out in Table 6.6. There are 30 quarters of data for the former but only 15 quarters for the emerging market (Latin American) equities. We can first calculate the truncated-sample estimator (Table 6.7). In this case, we only use the data from Q2 1997 to Q4 2000.

Figure 6.10 Scatter plot for overlapping data points



We see that the maximum likelihood results lead to an upward revision on the emerging market returns as these are highly correlated with the developed market returns (see Figure 6.10). Thus, part of the return differential between the full-sample and the short-sample periods will be added to the average return of the shorter series.

The most dramatic impact arises from the upward revision of the volatility estimate for emerging market equities, which is the direct effect of a much higher estimation error than for developed markets.

APPENDIX A: DERIVATION OF UNIVARIATE EXAMPLE

From Equation 6.10 we already know that

$$p(\mu \mid \boldsymbol{r}) \propto \exp\left\{-\frac{1}{2}\left(\frac{(\mu-\mu_{\text{prior}})^2}{\varphi^2}+\frac{\sum_{i=1}^{T}(r_i-\mu)^2}{\sigma^2}\right)\right\} \tag{6.61}$$

Expanding the term in the main set of inner brackets on the right-hand side, we get

$$\begin{aligned}
&\mu^2\left(\frac{1}{\varphi^2}+\frac{T}{\sigma^2}\right)-2\mu\left(\frac{T\hat{\mu}}{\sigma^2}+\frac{\mu_{\text{prior}}}{\varphi^2}\right)+\frac{\sum_{i=1}^{T}r_i^2}{\sigma^2}+\frac{\mu_{\text{prior}}^2}{\varphi^2}\\
&\quad=\left(\mu-\frac{T\hat{\mu}/\sigma^2+\mu_{\text{prior}}/\varphi^2}{1/\varphi^2+T/\sigma^2}\right)^2\\
&\qquad\times\left(\frac{1}{\varphi^2}+\frac{T}{\sigma^2}\right)+\frac{\sum_{i=1}^{T}r_i^2}{\sigma^2}+\frac{\mu_{\text{prior}}^2}{\varphi^2}\\
&\qquad-\frac{T\hat{\mu}/\sigma^2+\mu_{\text{prior}}/\varphi^2}{1/\varphi^2+T/\sigma^2}
\end{aligned} \tag{6.62}$$

It then follows that

$$\left.\begin{aligned} \mu_{\text{predictive}} &= \left(\frac{1}{\varphi_n^2}\mu_{\text{prior}} + \frac{T}{\sigma^2}\hat{\mu}\right)\left(\frac{1}{\varphi_T^2} + \frac{T}{\sigma^2}\right)^{-1} \\ \varphi_T &= \left(\frac{1}{\varphi^2} + \frac{T}{\sigma^2}\right)^{1/2} \end{aligned}\right\} \tag{6.63}$$

This technique is also known as completing the square.[37]

Instead of going through the analytics, we could also simulate the posterior distribution. We know that the posterior distribution can be written as $p(\boldsymbol{\theta} \mid \boldsymbol{r}) \propto p(\boldsymbol{r} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$. All we have to do is to simulate from the prior distribution $p(\boldsymbol{\theta})$ and substitute this draw in the likelihood function $p(\boldsymbol{r} \mid \boldsymbol{\theta})$. Multiplying both results and repeating these steps many times gives us the unnormalised posterior. We then add the realisations up and divide each realisation by the sum to get the normalised posterior.

APPENDIX B: ESTIMATION ERROR AND VOLATILITY FORECASTS

Risk arises from unpredictable variations; variations that can be forecast do not expose investors to risk. Suppose we employ a standard autoregressive approach, using dividend yields to forecast future equity returns as in the equation below[38]

$$\left.\begin{aligned} \begin{pmatrix} r_{t+1} \\ \mathrm{d}y_{t+1} \end{pmatrix} &= \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \mathrm{d}y_t + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix} \\ \boldsymbol{z}_{t+1} &= \boldsymbol{a} + \boldsymbol{b}x_t + \boldsymbol{\varepsilon}_t \end{aligned}\right\} \tag{6.64}$$

where returns are denoted by r and dividend yields are expressed as $\mathrm{d}y$, $\boldsymbol{z}_{t+1} = (r_{t+1}, \mathrm{d}y_{t+1})'$, $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \varepsilon_{2t})'$ and $x_t = \mathrm{d}y_t$. Equation 6.64 describes the relationship for a single time period, whereas we are interested in the average relationship over many periods. Rewriting this equation for all data points ("stacking") gives us

$$\underbrace{\begin{pmatrix} r_2 & \mathrm{d}y_2 \\ \vdots & \vdots \\ r_t & \mathrm{d}y_t \end{pmatrix}}_{\boldsymbol{Z}} = \underbrace{\begin{pmatrix} \mathbf{1} & \mathrm{d}y_1 \\ \vdots & \vdots \\ \mathbf{1} & \mathrm{d}y_{t-1} \end{pmatrix}}_{\boldsymbol{X}} \underbrace{\begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix}}_{\boldsymbol{C}} + \underbrace{\begin{pmatrix} \varepsilon_{12} & \varepsilon_{22} \\ \vdots & \vdots \\ \varepsilon_{1t} & \varepsilon_{2t} \end{pmatrix}}_{\boldsymbol{E}} \tag{6.65}$$

or in matrix form

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{C} + \boldsymbol{E} \tag{6.66}$$

Table 6.8 Monthly risk premium (RP) and dividend yield (DY) for UK equities

Quartal	RP	DY	Quartal	RP	DY	Quartal	RP	DY
Q1 75	−4.9	9.9	Q4 83	−1.5	5.2	Q3 92	−5.1	5.0
Q2 75	73.2	5.9	Q1 84	15.2	4.6	Q4 92	3.1	4.8
Q3 75	8.7	5.5	Q2 84	6.5	4.6	Q1 93	11.0	4.4
Q4 75	9.5	5.2	Q3 84	−8.9	5.2	Q2 93	4.4	4.0
Q1 76	12.3	4.7	Q4 84	13.9	4.9	Q3 93	1.7	4.0
Q2 76	0.0	4.8	Q1 85	16.3	4.4	Q4 93	11.1	3.7
Q3 76	−1.7	5.1	Q2 85	2.9	4.6	Q1 94	10.8	3.3
Q4 76	−20.8	6.8	Q3 85	−3.9	4.9	Q2 94	−5.2	3.6
Q1 77	26.2	5.5	Q4 85	7.1	4.7	Q3 94	−3.4	3.9
Q2 77	11.0	5.2	Q1 86	4.7	4.6	Q4 94	1.2	3.9
Q3 77	8.4	5.1	Q2 86	20.3	3.9	Q1 95	−0.6	4.0
Q4 77	12.0	4.7	Q3 86	−3.1	4.2	Q2 95	3.3	4.0
Q1 78	−5.2	5.2	Q4 86	1.3	4.2	Q3 95	8.0	3.8
Q2 78	−3.4	5.5	Q1 87	12.7	3.9	Q4 95	3.5	3.8
Q3 78	11.3	5.1	Q2 87	9.5	3.6	Q1 96	4.5	3.8
Q4 78	4.2	5.1	Q3 87	27.5	3.0	Q2 96	5.4	3.7
Q1 79	0.4	5.2	Q4 87	−5.7	3.3	Q3 96	−4.0	4.0
Q2 79	19.9	4.6	Q1 88	−22.7	4.3	Q4 96	10.4	3.7
Q3 79	−8.0	5.8	Q2 88	0.4	4.4	Q1 97	5.2	3.6
Q4 79	5.0	5.8	Q3 88	5.1	4.3	Q2 97	2.1	3.6
Q1 80	−2.3	6.1	Q4 88	−1.2	4.5	Q3 97	12.7	3.3
Q2 80	1.7	6.7	Q1 89	2.4	4.6	Q4 97	7.7	3.1
Q3 80	15.1	6.0	Q2 89	10.7	4.4	Q1 98	−0.3	3.1
Q4 80	5.3	5.8	Q3 89	10.7	4.1	Q2 98	14.4	2.7
Q1 81	−3.5	6.1	Q4 89	−7.5	4.7	Q3 98	2.6	2.7
Q2 81	12.5	5.5	Q1 90	8.9	4.5	Q4 98	−17.0	3.2
Q3 81	−0.7	5.7	Q2 90	−5.7	5.0	Q1 99	14.7	2.9
Q4 81	−10.8	6.6	Q3 90	8.0	4.7	Q2 99	10.4	2.4
Q1 82	10.2	6.1	Q4 90	−13.4	5.6	Q3 99	3.1	2.4
Q2 82	1.9	6.2	Q1 91	−1.3	5.7	Q4 99	−9.5	2.7
Q3 82	5.2	6.1	Q2 91	24.1	4.7	Q1 00	13.7	2.5
Q4 82	15.6	5.4	Q3 91	1.0	4.8	Q2 00	−6.4	2.5
Q1 83	3.7	5.3	Q4 91	2.2	4.8	Q3 00	5.1	2.3
Q2 83	12.8	4.8	Q1 92	−1.7	4.9	Q4 00	−2.5	2.4
Q3 83	−1.0	5.0	Q2 92	5.3	4.7	Q1 01	−2.2	2.5

Source: Datastream.

We recognise this as a simple regression model that can be estimated using ordinary least squares (OLS), giving the well-known solution[39]

$$\hat{\boldsymbol{C}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z} \tag{6.67}$$

The 2×2 covariance matrix of error terms (regression residuals) can be calculated in the usual way

$$\hat{\boldsymbol{\Sigma}} = \frac{(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{C}})'(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{C}})}{T} = \frac{\hat{\boldsymbol{E}}'\hat{\boldsymbol{E}}}{T} \tag{6.68}$$

where T is the number of observations.

We will use sample data from Table 6.8 to estimate our simple model. The matrix of regression coefficients shows the high persistency of dividend yields (shown by an autocorrelation of 0.79)

$$\hat{\boldsymbol{C}} = \begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix} = \begin{pmatrix} -14.89 & 0.87 \\ 4.3 & 0.79 \end{pmatrix} \tag{6.69}$$

while the covariance matrix of regression residuals exhibits a negative correlation between shocks to the system in Equation 6.64

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \operatorname{cov}(\hat{\varepsilon}_1, \hat{\varepsilon}_1) & \operatorname{cov}(\hat{\varepsilon}_1, \hat{\varepsilon}_2) \\ \operatorname{cov}(\hat{\varepsilon}_2, \hat{\varepsilon}_1) & \operatorname{cov}(\hat{\varepsilon}_2, \hat{\varepsilon}_2) \end{pmatrix} = \begin{pmatrix} 100.01 & -4.9 \\ -4.9 & 0.27 \end{pmatrix} \tag{6.70}$$

The correlation between ε_1 and ε_2 amounts to -0.94 and can be calculated from

$$-4.9 \times \frac{1}{\sqrt{100.01}} \times \frac{1}{\sqrt{0.27}}$$

Our simple regression model contains all information on the time-variable investment opportunities that are available. Out-of-sample forecasts (in per cent) for the next two quarters (Q2 and Q3) in 2001 for the UK equity market are

$$\begin{aligned} r_{\mathrm{Q2}} &= -14.89 + 4.3 \cdot 2.5 = -4.14 \\ r_{\mathrm{Q3}} &= -14.89 + 4.3 \cdot (0.87 + 0.79 \cdot 2.5) = -2.66 \end{aligned}$$

We are now in a position to generate forecasts for returns in period $t + n$

$$\boldsymbol{z}_{t+1} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{z}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{B} = \begin{bmatrix} 0 & b_1 \\ 0 & b_2 \end{bmatrix}, \quad \boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \quad \boldsymbol{\varepsilon}_t = \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} \tag{6.71}$$

We will start with a simple two-period example. Knowing the return-generating process for $t + 1$ and $t + 2$

$$\boldsymbol{z}_{t+1} = \hat{\boldsymbol{a}} + \hat{\boldsymbol{B}}\boldsymbol{z}_t + \boldsymbol{\varepsilon}_t \tag{6.72}$$

$$\boldsymbol{z}_{t+2} = \hat{\boldsymbol{a}} + \hat{\boldsymbol{B}}\boldsymbol{z}_{t+1} + \boldsymbol{\varepsilon}_{t+1} \tag{6.73}$$

we can use substitution from Equations 6.72 and 6.73 to generate a two-period forecast

$$\begin{aligned} z_{t+2} &= \hat{\boldsymbol{a}} + \hat{\boldsymbol{B}} z_{t+1} + \boldsymbol{\varepsilon}_{t+1} \\ &= \hat{\boldsymbol{a}} + \hat{\boldsymbol{B}} \underbrace{(\hat{\boldsymbol{a}} + \hat{\boldsymbol{B}} z_t + \boldsymbol{\varepsilon}_t)}_{z_{t+1}} + \boldsymbol{\varepsilon}_{t+1} \\ &= \hat{\boldsymbol{a}} + \hat{\boldsymbol{a}}\hat{\boldsymbol{B}} + \hat{\boldsymbol{B}}^2 z_t + \hat{\boldsymbol{B}}\boldsymbol{\varepsilon}_t + \boldsymbol{\varepsilon}_{t+1} \end{aligned}$$

Generalisation yields the return for period $t + n$

$$\begin{aligned} z_{t+n} = {} & \hat{\boldsymbol{a}} + \hat{\boldsymbol{B}}\hat{\boldsymbol{a}} + \hat{\boldsymbol{B}}^2\hat{\boldsymbol{a}} + \cdots \\ & + \hat{\boldsymbol{B}}^{n-1}\hat{\boldsymbol{a}} + \hat{\boldsymbol{B}}^{n} z_t + \boldsymbol{\varepsilon}_{t+n} + \hat{\boldsymbol{B}}\boldsymbol{\varepsilon}_{t+(n-1)} + \cdots \\ & + \hat{\boldsymbol{B}}^{n-1}\boldsymbol{\varepsilon}_{t+1} \end{aligned} \tag{6.74}$$

Equation 6.74 allows us to forecast any point in the future; note that, as the expected value of a residual is zero, all residual terms drop out – ie, become zero – if we take expectations. In the case of no forecasting power (where $\boldsymbol{B} = \boldsymbol{0}$), our forecast coincides with the unconditional mean ($\hat{a}_1$ in $\hat{\boldsymbol{a}}$). The remaining step is to write down the conditional forecast (conditional on information known at t) and calculate the covariance matrix for the n-period returns $(z_{t+1} + \cdots + z_{t+n})$. This is done in the following equation

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}(n) = {} & \hat{\boldsymbol{\Sigma}} + (\boldsymbol{I} + \hat{\boldsymbol{B}})\hat{\boldsymbol{\Sigma}}(\boldsymbol{I} + \hat{\boldsymbol{B}})' + (\boldsymbol{I} + \hat{\boldsymbol{B}} + \hat{\boldsymbol{B}}^2)\hat{\boldsymbol{\Sigma}}(\boldsymbol{I} + \hat{\boldsymbol{B}} + \hat{\boldsymbol{B}}^2) + \cdots \\ & + (\boldsymbol{I} + \hat{\boldsymbol{B}} + \hat{\boldsymbol{B}}^2 + \cdots + \hat{\boldsymbol{B}}^{n-1})\hat{\boldsymbol{\Sigma}}(\boldsymbol{I} + \hat{\boldsymbol{B}} + \hat{\boldsymbol{B}}^2 + \cdots + \hat{\boldsymbol{B}}^{n-1})' \end{aligned} \tag{6.75}$$

Admittedly this formula looks very unfriendly, but we will justify it using a two-period example. We first apply the variance operator to

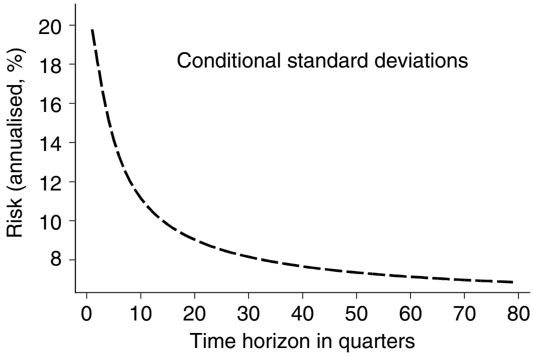
$$z_{t+1} + z_{t+2} = \underbrace{\hat{\boldsymbol{a}} + \hat{\boldsymbol{B}} z_t + \boldsymbol{\varepsilon}_t}_{z_{t+1}} + \underbrace{\hat{\boldsymbol{a}} + \hat{\boldsymbol{a}}\hat{\boldsymbol{B}} + \hat{\boldsymbol{B}}^2 z_t + \hat{\boldsymbol{B}}\boldsymbol{\varepsilon}_t + \boldsymbol{\varepsilon}_{t+1}}_{z_{t+2}}$$

However, uncertainty arises only from the residual terms as the model parameters are assumed to be fully known in t (no estimation error). We then get

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}(\mathbf{2}) &= \operatorname{var}(z_{t+1} + z_{t+2}) = \operatorname{var}(\boldsymbol{\varepsilon}_t + \hat{\boldsymbol{B}}\boldsymbol{\varepsilon}_t + \boldsymbol{\varepsilon}_{t+1}) \\ &= \operatorname{var}(\boldsymbol{\varepsilon}_t(\boldsymbol{I} + \hat{\boldsymbol{B}}) + \boldsymbol{\varepsilon}_{t+1}) = (\boldsymbol{I} + \hat{\boldsymbol{B}})\hat{\boldsymbol{\Sigma}}(\boldsymbol{I} + \hat{\boldsymbol{B}})' + \hat{\boldsymbol{\Sigma}} \end{aligned} \tag{6.76}$$

We assume here that the residual terms are stationary (covariance in period two is the same as in period one) and uncorrelated across time. For a better understanding of Equation 6.76, we can again

Figure 6.11 Equity risk and time horizon



investigate the non-information case ($\hat{\boldsymbol{B}} = \mathbf{0}$). In this circumstance, we get the familiar "square root of time" rule

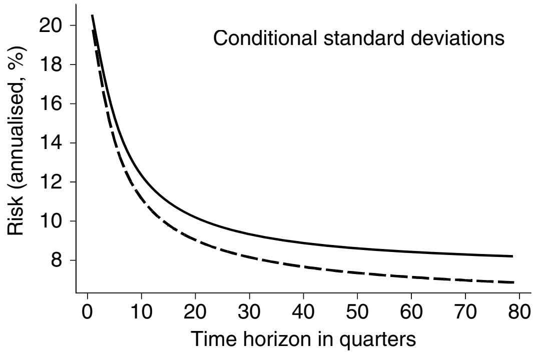
$$\hat{\boldsymbol{\Sigma}}(n) = \hat{\boldsymbol{\Sigma}} + \cdots + \hat{\boldsymbol{\Sigma}} = n\hat{\boldsymbol{\Sigma}}$$

Decomposing Equation 6.76 further, we get

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}(\mathbf{2}) &= \hat{\boldsymbol{\Sigma}} + (\boldsymbol{I} + \hat{\boldsymbol{B}})\hat{\boldsymbol{\Sigma}}(\boldsymbol{I} + \hat{\boldsymbol{B}})' \\ &= \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \\ &\quad + \left(\begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & b_1 \\ \mathbf{0} & b_2 \end{bmatrix} \right) \\ &\qquad \times \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \left(\begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & b_1 \\ \mathbf{0} & b_2 \end{bmatrix} \right)' \\ &= \begin{bmatrix} 2\sigma_{11} + 2b_1\sigma_{12} + b_1^2\sigma_{22} & \cdots \\ \cdots & \cdots \end{bmatrix}_{2\times 2} \end{aligned} \tag{6.77}$$

Only in the case of no forecasting opportunities (where $b = 0$) will the variance of the two-period returns equal twice the variance of the one-period returns. If $2b_1\sigma_{12} + b_1^2\sigma_{22} < 0$, we see that the conditional variance of the two-period returns is less than twice the variance of one-period returns. Substituting our estimated parameters, it is easy to see that these conditions are satisfied. The important term is the covariance between the residuals in Equation 6.64. If they are sufficiently negative, shocks on dividend yields will be compensated by shocks on returns with the opposite sign. The more pronounced the

Figure 6.12 Risk, time horizon and estimation error



impact of dividend yields (b_1) on tomorrow's returns, the stronger this effect will be.

Figure 6.11 plots the relationship between time horizon and conditional volatility (annualised) for up to 80 quarters (20 years). This is done using Equation 6.75 for varying horizons. Time-variable investment opportunities (good when dividend yields are high and bad when they are low) result in good years followed by bad years and, hence, to a decrease in the unexplained volatility. Remember, it is only unexplained volatility (unexplained with respect to information at t) that exposes investors to risk; movements that can be anticipated are not a source of risk.

So far, we have accepted point forecasts from OLS as 100% true; uncertainty arose from the volatility in the residual terms, not from estimation error. However, we know that our model forecasts are affected by uncertainty.

To reflect estimation error in our return-generating model (B1), we will use the procedure already described in Section 6.1. Here the parameters of concern are $\boldsymbol{\theta} = (\boldsymbol{C}, \boldsymbol{\Sigma})$. Results are given in Figure 6.12.

EXERCISES

1. Download a series of US dollar/euro exchange rates. Recall that value-at-risk violations follow a binomial distribution (1 for a return outside the 95% confidence level with 5% probability and 0 with 95% probability). Assume a beta prior on value-at-risk violations and plot the posterior distribution for various priors.

2. You are given two ($i = 1, 2$) independent sources of information on the "alphas" of assets in a given investment universe. Each source offers you mean and variance α_i, σ_i of the unknown α. How should you weight these two sources of information?

3. Suppose we have five years of annual historical data on equities and bonds with

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} 10 \\ 7 \\ 4 \end{bmatrix} \quad \text{and} \quad \hat{\boldsymbol{\Omega}} = \begin{bmatrix} 400 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 25 \end{bmatrix}$$

The investor's risk aversion is $\lambda = 0.075$. They have an extremely uninformative prior. The cash rate is 2%. How does the efficient frontier change, when we introduce estimation error? Does the efficient set change? How will the maximum Sharpe ratio change? What is the only asset that exhibits neither investment nor estimation risk?

4. Assume a market with 10 assets and equal correlation ($\rho = 0.8$) and equal volatility each ($\sigma = 20\%$). The equilibrium return on each is 5%. Assume $\tau = 1$.

(a) What is the equilibrium return on the market?

(b) You only forecast asset 1 with $q = 10\%$. Assume the analyst expects their forecast to range between +20% and 0% with 66% probability. How does the forecast for asset 1 change? How do the forecasts for all other assets change?

(c) How do the results in (b) change for $\rho = 0.2$?

(d) How does (b) change if we use $\tau = 0.1$ or $\tau = 10$?

[1] Markowitz (1987, p. 57).

[2] It has been argued that this judgmental part of traditional statistics already uses non-sample information and therefore that even frequentists implicitly use prior information.

[3] $p(\boldsymbol{r}, \boldsymbol{\theta}) = p(\boldsymbol{r} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$ means that the probability of getting $\boldsymbol{r}$ and $\boldsymbol{\theta}$ is the same as drawing $\boldsymbol{r}$ conditional on a particular realisation on $\boldsymbol{\theta}$ times the probability of arriving at this realisation of $\boldsymbol{\theta}$ called $p(\boldsymbol{\theta})$.

[4] See Klein and Bawa (1976) for an application in a two-asset context.

[5] See Green (2000) or Borse (1997) for a discussion of Monte Carlo integration.

[6] The example is taken from Barberis (2000).

[7] See Kritzman (2000) for an overview of the time-diversification arguments.

[8] The estimation error on the mean return is given by $\sigma / \sqrt{T}$, where T denotes the number of observations. Figure 6.4 plots the distribution of estimated mean returns for T equal to 60, 120 and 240.

[9] Conjugate priors have the same functional form (belong to the same family of distributions) as the likelihood function. This class of priors will be used throughout this chapter. See Morris (1983) for the exponential family of distributions. Conjugate priors are used for mathematical convenience as well as giving the prior a sample interpretation.

[10] See Verbeek (2000, Chapter 7).

[11] See Theil (1971). A prerequisite of this interpretation is that we use a conjugate prior.

[12] We use an example similar to that in the book by Rudd and Clasing (1988).

[13] See Grinold and Kahn (2000) for a review of the literature.

[14] See Sharpe (1991) for an exposition on the arithmetic of active management.

[15] This section follows O'Cinneide (2001). Other expositions using a similar set up can be found in Frost and Savarino (1986) or Jorion (1986).

[16] Turner and Hensel (1993) found that average returns within asset groups (bonds, equities, cash) were statistically undistinguishable between countries. Deriving priors of this form would be called empirical Bayes.

[17] The parameters of the prior distribution are also called hyper-parameters. They are assumed to be known with no uncertainty.

[18] Note that this does not mean it is normal. In fact, it has been shown by Barry (1974) that if there is uncertainty about both mean returns and the covariance matrix, the predictive distribution is *t*-distributed. Technically, this would require the investor to possess quadratic utility in order to rescue mean–variance based portfolio selection.

[19] Note that Equation 6.14 is very similar to Equation 6.11 where precision is now captured by the inverse on a variance–covariance matrix. Since $\boldsymbol{\Psi}_{\text{prior}}^{-1}$ contains estimation errors rather than asset variances, the elements in $\hat{\boldsymbol{\Omega}}^{-1}$ have to be adjusted to arrive at the same order of magnitude. Note that $T\hat{\boldsymbol{\Omega}}^{-1} = (1/T)^{-1}\hat{\boldsymbol{\Omega}}^{-1}$, which corrects for different orders of magnitude.

[20] See Jeffreys (1961).

[21] See also Barry (1974) or Frost and Savarino (1986).

[22] See Frost and Savarino (1986, p. 296).

[23] A slight modification of this is the case of uninformative priors, where return series exhibit different lengths. Stambaugh (1997) shows that not only is estimation error higher for shorter series than for longer ones, but additionally that means change as long as assets are not completely uncorrelated. The mean difference between the overlapping parts of two series is pulled to the mean of a similar but longer return series (a correlated series with less estimation error).

[24] We have already shown that non-informative priors do not change the set of efficient solutions.

[25] The model described below is the Black–Litterman model (Black and Litterman 1992). An excellent review can be found in Satchell and Scowcroft (1998) or Lee (2000).

[26] See Black and Litterman (1992, p. 43, Equation 8).

[27] See the experience report from Bevan and Winkelmann (1998).

[28] A review of the CAPM can be found in Cuthbertson (1996, Chapter 3).

[29] See Pastor (2000) for more on this.

[30] This is also the case for multiple managers. See Chapter 11 for a more detailed discussion.

[31] $\beta_i = \operatorname{cov}(r_i, \sum_{i=1}^{k} w_i r_i)\operatorname{var}(\sum_{i=1}^{k} w_i r_i)^{-1}$.

[32] See Scherer and Martin (2005) for a worked Bayesian regression example in S^{+}Bayes.

[33] Longer series are available for the US from Ibbotson Associates, Chicago. Dimson *et al* (2001) recovered historical time series for all major stock and bond markets for the last 101 years.

[34] The reader is referred to Stambaugh (1997) for a complete derivation of the methodology borrowed above.

[35] Reassuringly, the maximum likelihood estimator ensures that the covariance matrix of returns is still positive definite. Other estimators – such as the use of long history on long series and short history on short series, including co-movements between long and short series – do not necessarily have these characteristics.

[36] See Chiang (1984, p. 42).

[37] See, for example, Cochrane (2001), Fama and French (1988), Shiller (1996) or Goetzmann and Jorion (1993) on this approach.

[38] Readers are referred to Lütkepohl (1991) as the standard reference on vector autoregressive modelling.

[39] See Comon (2001) for a similar study on emerging market equities.

REFERENCES

Barberis, N., 2000, "Investing for the Long Run when Returns Are Predictable", *Journal of Finance* 55, pp. 225–64.

Barry, C., 1974, "Portfolio Analysis under Uncertain Means", *Journal of Finance* 29, pp. 515–22.

Bevan, A., and K. Winkelmann, 1998, "Using the Black–Litterman Global Asset Allocation Model: Three Years of Experience", Working Paper, Goldman Sachs.

Black, F., and R. Litterman, 1992, "Global Portfolio Optimization", *Financial Analysts Journal* 48, pp. 28–43.

Borse, G., 1997, *Numerical Methods with Matlab* (Boston: Thomson).

Cheung, W., 2009, "Efficient Bayesian Factor Mimicking: Methodology, Tests and Comparison", Available at SSRN: http://ssrn.com/abstract=1457022.

Chiang, A., 1984, *Fundamental Methods of Mathematical Economics*, Third Edition (New Jersey: McGraw-Hill).

Cochrane, J. H., 2001, *Asset Pricing* (Princeton, NJ: Princeton University Press).

Comon, E., 2001, "Essays on Investment and Consumption Choice", PhD Thesis, Harvard University.

Cuthbertson, K., 1996, *Quantitative Financial Economics* (Chichester: John Wiley & Sons).

Dimson, E., P. Marsh and M. Staunton, 2001, *Millennium Book II: 101 Years of Investment Returns* (London: ABN AMRO).

Fama, E. F., and K. R. French, 1988, "Dividend Yields and Expected Stock Returns", *Journal of Financial Economics* 24, pp. 23–49.

Frost, P., and J. Savarino, 1986, "An Empirical Bayes Approach to Efficient Portfolio Selection", *Journal of Financial and Quantitative Analysis* 21, pp. 293–305.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin, 2003, *Bayesian Data Analysis*, Second Edition (Chapman & Hall).

Goetzmann, W., and P. Jorion, 1993, "Testing the Predictive Power of Dividend Yields", *Journal of Finance* 48, pp. 663–79.

Green, W., 2000, *Econometric Analysis*, Fourth Edition (Englewood Cliffs, NJ: Prentice Hall).

Grinold, R., and R. Kahn, 2000, *Active Portfolio Management*, Second Edition (New York: McGraw-Hill).

Jeffreys, H., 1961, *Theory of Probability*, Third Edition (Oxford University Press).

Jorion, P., 1986, "Bayes–Stein Estimation for Portfolio Analysis", *Journal of Financial and Quantitative Analysis* 21, pp. 279–92.

Klein, R., and S. Bawa, 1976, "The Effect of Estimation Risk on Optimal Portfolio Choice", *Journal of Financial Economics* 3, pp. 215–31.

Kritzman, M., 2000, *Puzzles of Finance* (New York: John Wiley & Sons).

Lee, W., 2000, *Theory and Methodology of Tactical Asset Allocation*, Frank J. Fabozzi Series (New York: John Wiley & Sons).

Lütkepohl, H., 1991, *Introduction to Multiple Time Series Analysis* (Heidelberg: Springer).

Markowitz, H., 1987, *Mean–Variance Analysis in Portfolio Choice and Capital Markets* (Blackwell).

Morris, C., 1983, "Natural Exponential Families with Quadratic Variance Functions: Statistical Theory", *Annals of Statistics* 11, pp. 519–29.

O'Cinneide, C., 2001, *A Bayesian Approach to Portfolio Selection* (New York: Deutsche Asset Management).

Pastor, L., 2000, "Portfolio Selection and Asset Pricing Models", *Journal of Finance* 55, pp. 179–221.

Rudd, A., and H. K. Clasing Jr., 1988, *Modern Portfolio Theory: The Principles of Investment Management* (Irwine: Dow Jones).

Satchell, S., and A. Scowcroft, 1998, "A Demystification of the Black–Litterman Model, Managing Quantitative and Traditional Portfolio Construction", Working Paper, Judge Institute.

Scherer, B., and D. Martin, 2005, *Introduction to Modern Portfolio Optimization with NuOPT™, S-Plus®, and S+ Bayes™* (New York: Springer).

Sharpe, W., 1991, "The Arithmetic of Active Management", *Financial Analysts Journal* 47(1), pp. 7–9.

Shiller, R. J., 1996, "Price-Earnings Ratios as Forecasters of Returns: The Stock Market Outlook in 1996", Working Paper, Yale University.

Stambaugh, R., 1997, "Analysing Investments Whose Histories Differ in Length", *Journal of Financial Economics* 45, pp. 285–331.

Theil, H., 1971, *Principles of Econometrics* (Amsterdam: John Wiley & Sons).

Turner, A., and C. Hensel, 1993, "Were the Returns from Stocks and Bonds of Different Countries Really Different in the 1980s?", *Management Science* July, pp. 835–44.

Verbeek, M., 2000, *A Guide to Modern Econometrics* (Chichester: John Wiley & Sons).

Zellner, A., 1971, *An Introduction to Bayesian Inference in Econometrics* (New York: John Wiley & Sons).

7

Testing Portfolio Construction Methodologies Out-of-Sample

7.1 INTRODUCTION

Various portfolio optimisation and data refinement techniques have been proposed in the academic literature to address estimation error and its impact on portfolio construction.[1] While the academic literature remains firmly rooted in Bayesian decision theory, Michaud (1998) deviated from this tradition and suggested a portfolio construction heuristic[2] using resampling techniques (called resampled efficiency or RE),[3] based on the earlier work by Jorion (1992).[4] Even though, the in-sample properties of RE have already been subject to substantial critique,[5] little attempt has been made to compare the out-of-sample performance for portfolio resampling with more simple and established methodologies. Markowitz and Usmen (2003) are an exemption. Their careful study reveals that RE is superior to both the Bayesian alternative as well as to Markowitz's original method.[6] However, the outcomes of their study rely on a particular data set as well as on a specific Bayesian prior. Moreover, they did not allow for the possibility of short sales. That is why they titled their research: an experiment.

The purpose of this section is to provide another experiment that also provides two additional insights into the out-of-sample performance of RE. First, we extend our analysis to the relative performance for an unconstrained investor (hedge funds, active managers with small tracking errors, ie, non-binding long constraints, etc). We regard this case as particularly interesting because estimation error has its biggest impact on performance in this setting. Any methodology that aims at handling estimation error should be able to work in this case, where we need it most. Second, we contrast Markowitz and Usmen with the use of a James and Stein prior, rather than a diffuse prior. Our choice is based on two arguments. We find overwhelming evidence for the usefulness of this prior for an unconstrained manager, which makes it a natural choice for the constrained case. Further, we find that resampling only differs from Markowitz optimisation, if portfolios are constrained (see Section 7.1). At the same time recent research by Jagannathan and Ma (2003) has shown that imposing position limits provides a shrinkage estimator. Putting both observations together we reason that the out-of-sample success of RE is likely to be rooted in some sort of (uncontrolled) shrinkage.

Using our own experimental setup we confirm that Michaud's method outperforms naive Markowitz optimisation in many circumstances (however, only for a long-only constraint). Contrary to Markowitz and Usmen (2003) we, however, find that James–Stein shrinkage outperforms Michaud's method. Even though we somehow engineered this result, it proves a more general point:

> Out-of-sample testing of Bayesian methods is difficult in concept. For every distribution we will find a prior that will outperform resampling (and vice versa). Equally for every prior, we will find a distribution where resampling outperforms. The fact that one method outperforms another for a given set of data means little. In the absence of theory, investors don't know when one method will outperform the other as they don't know the true distribution.

7.2 RESAMPLED EFFICIENCY

Instead of running a single portfolio optimisation with a given set of sample estimates, Michaud (1998) suggests running k optimisations using resampled inputs. Optimal portfolios are then found by averaging across these k allocations.[7] Mathematically we find efficient allocations ($\boldsymbol{w}^*_{\text{res}}$) by using[8]

$$\boldsymbol{w}^*_{\text{res}} = \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{w}^*_{i,\text{res}} = \frac{1}{k}\sum_{i=1}^{k} \arg\max_{w \in C}(\boldsymbol{w}^{\mathrm{T}}\bar{\boldsymbol{\mu}}_{i,\text{res}} - \tfrac{1}{2}\lambda \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Omega}_0 \boldsymbol{w}) \qquad (7.1)$$

where $\bar{\boldsymbol{\mu}}_{i,\text{res}}$ denotes the vector of sample mean estimates and $\boldsymbol{w}^*_{i,\text{res}}$ its associated optimal solution for the ith resampling. Constraints are subsumed in the notation $\boldsymbol{w} \in \boldsymbol{C}$. We also assumed that the covariance matrix ($\boldsymbol{\Omega}_0$) of returns is known, ie, each resampling draws the vector of expected returns from

$$\bar{\boldsymbol{\mu}}_{i,\text{res}} \sim N(\bar{\boldsymbol{\mu}}, \tau^{-1}\boldsymbol{\Omega}_0) \qquad (7.2)$$

The number of draws (τ) per resampling can be heuristically thought of as the level of confidence we have in a given data set.[9] If the number of draws is very high the dispersion in mean estimates is low, while if it is low the variation in resampled means will be high. We will now establish some properties of Equation 7.1. In the absence of long-only constraints we can find the optimal solution to the *i*th draw as

$$\boldsymbol{w}_{i,\mathrm{res}}^{*} = \arg\max(\boldsymbol{w}^{\mathrm{T}}\bar{\boldsymbol{\mu}}_{i,\mathrm{res}} - \tfrac{1}{2}\lambda\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Omega}_0\boldsymbol{w}) = \lambda^{-1}\boldsymbol{\Omega}_0^{-1}\bar{\boldsymbol{\mu}}_{i,\mathrm{res}} \tag{7.3}$$

Averaging across *k* draws leads to

$$\boldsymbol{w}_{\mathrm{res}}^{*} = \frac{1}{k}\sum_{i=1}^{k}\boldsymbol{w}_{i,\mathrm{res}}^{*} = \lambda^{-1}\boldsymbol{\Omega}_0^{-1}\left(\frac{1}{k}\sum_{i=1}^{k}\bar{\boldsymbol{\mu}}_{i,\mathrm{res}}\right)\underset{k\to\infty}{\approx}\lambda^{-1}\boldsymbol{\Omega}_0^{-1}\bar{\boldsymbol{\mu}} = \boldsymbol{w}_{\mathrm{mv}}^{*} \tag{7.4}$$

As we can decompose $\bar{\boldsymbol{\mu}}_{i,\mathrm{res}} = \bar{\boldsymbol{\mu}} + \boldsymbol{\varepsilon}_{i,\mathrm{res}}$, we arrive at Equation 7.4 since

$$\lim_{k\to\infty}\lambda^{-1}\boldsymbol{\Omega}_0^{-1}\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{\varepsilon}_{i,\mathrm{res}} = 0$$

Resampling effectively becomes the traditional Markowitz solution distribution plus noise from resampling.[10] Note that the noise term will eventually converge to zero for larger *k* (or τ). We can also use the above decomposition to calculate the distribution of resampled weights from

$$\begin{aligned}\boldsymbol{\Omega}_{\boldsymbol{w}_{\mathrm{res}}^{*}} &= \mathrm{cov}(\lambda^{-1}\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{\mu}}_{i,\mathrm{res}}) = \lambda^{-2}\boldsymbol{\Omega}^{-1}\,\mathrm{cov}(\bar{\boldsymbol{\mu}}_{i,\mathrm{res}})\boldsymbol{\Omega}^{-1}\\ &= \lambda^{-2}\boldsymbol{\Omega}^{-1}\frac{\boldsymbol{\Omega}}{\tau}\boldsymbol{\Omega}^{-1} = \lambda^{-2}\tau^{-1}\boldsymbol{\Omega}^{-1}\end{aligned} \tag{7.5}$$

For unconstrained optimisations, we have seen that RE and Markowitz coincide. Portfolios constructed using RE will not pick up the effect of estimation error on portfolio optimisation at all. Portfolio risk stays the same, even though the world became a riskier place (investors face investment and estimation risk). This is in stark contrast to Bayesian methods as well as robust optimisation.[11] It suggests that we can get rid of the estimation error problem by allowing investors to go short. This of course is nonsense. Whatever goes on with resampling has more to do with the effects of constraints on portfolio construction, rather than with addressing estimation error in the first place. Estimation error has its largest impact on portfolios, if left unconstrained and yet in this case (where we would need it most), RE fails.

A second important property is the fact that RE does not add extra sample information. In fact it uses resampling to derive better estimates which is somewhat in contrast to the statistical literature, where resampling is used foremost to come up with confidence bands. Also note, that resampling does not address estimation error in the first place as it effectively suffers from estimation error heritage. Repeatedly drawing from data that are measured with error, will simply transfer this error to the resampled data.

7.3 COMPETING METHODOLOGIES AND MODEL SET UP

We want to compare now the out-of-sample performance of RE versus naive Markowitz optimisation as well as a Bayesian alternative. Let us briefly introduce the employed methodologies. For simplicity we continue with our assumption that uncertainty only arises with mean estimates. Covariances are assumed to be known with certainty. The traditional approach is to use the maximum likelihood estimates (sample means) as inputs into Markowitz portfolio optimisation. Optimal weights are given by

$$\boldsymbol{w}_{\mathrm{mv}}^{*} = \arg\max_{\boldsymbol{w}\in\boldsymbol{C}}(\boldsymbol{w}^{\mathrm{T}}\bar{\boldsymbol{\mu}} - \tfrac{1}{2}\lambda\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Omega}_0\boldsymbol{w}) \tag{7.6}$$

where again risk aversion is given by λ, and $\bar{\boldsymbol{\mu}}$ denotes the (maximum likelihood) sample estimates of the unknown true means $\boldsymbol{\mu}_0$. The known covariance structure of asset returns is described by $\boldsymbol{\Omega}_0$. We label the results of this approach with “mv”.

The second approach is to replace the maximum likelihood estimates with James and Stein (1961) estimates. Stein estimates ($\boldsymbol{\mu}_{\mathrm{st}}$) place weights to historical data and prior knowledge depending on how close the sample estimates are to the used prior ($\boldsymbol{\mu}_{\mathrm{prior}}$). If the sample data conflict strongly with the prior, more weight is given to the data. The shrinkage estimator can be found in the following equation

$$\boldsymbol{\mu}_{\mathrm{st}} = (1-\theta)\hat{\boldsymbol{\mu}} + \theta\boldsymbol{\mu}_{\mathrm{prior}} \tag{7.7}$$

Note that defining $\boldsymbol{d} = \bar{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\mathrm{prior}}$ we use

$$\boldsymbol{\theta} = \min\left(\frac{n-2}{n_{\mathrm{sample}} - n - 2}\frac{1}{\boldsymbol{d}^{\mathrm{T}}\boldsymbol{\Omega}_0^{-1}\boldsymbol{d}'}, 1\right)$$

where n_{sample} denotes the number of sample date points and n describes the number of assets. We use the grand mean (mean of all observations across all assets) of the sample data for $\boldsymbol{\mu}_{\text{prior}}$, ie, all assets are expected to earn the same return. The statistical distance $\boldsymbol{d}^{\mathrm{T}}\boldsymbol{\Omega}_0^{-1}\boldsymbol{d}$ is used as a measure of "closeness". If the sample means are close to our prior the statistical distance is small and we put a large weight to our prior. Effectively this forces the optimal allocations towards the minimum variance portfolio. This is the only portfolio free of estimation risk (remember we assumed covariances to be known with certainty).[12] We use $\boldsymbol{\mu}_{\text{st}}$ instead of $\boldsymbol{\mu}$ in an ordinary Markowitz optimisation

$$\boldsymbol{w}_{\text{st}}^* = \underset{w \in C}{\arg\max}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_{\text{st}} - \tfrac{1}{2}\lambda\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Omega}_0\boldsymbol{w}) \tag{7.8}$$

All three methodologies use Markowitz optimisation as their optimisation methodology.

Out-of-sample testing those three portfolio construction methodologies involves more than just calculating performance along a historical return path. Nothing guarantees a given historical path to be representative. We therefore need to employ Monte Carlo simulation. The test procedure used here can be split into several steps.

Step 1. Choose a true covariance matrix $(\boldsymbol{\Omega}_0)$ and mean vector $(\boldsymbol{\mu}_0)$.

Step 2. Draw random mean returns (assuming covariances to be known with certainty) according to $\bar{\boldsymbol{\mu}}_i \sim N(\boldsymbol{\mu}_0, (1/\tau)\boldsymbol{\Omega}_0)$, where τ reflects the degree of estimation risk around $\boldsymbol{\mu}_0$. If we assume monthly data τ, will denote the number of (non-overlapping) monthly returns.

Step 3. Calculate optimal portfolios across all m methodologies, where $m \in (\text{mv}, \text{st}, \text{res})$. Each methodology only knows $\bar{\boldsymbol{\mu}}_i$ from Step 2. The Michaud algorithm resamples $i = 1, \dots, k$ mean vectors and the associated optimal portfolios from $\bar{\boldsymbol{\mu}}_i \sim N(\bar{\boldsymbol{\mu}}_i \tau^{-1}\boldsymbol{\Omega}_0)$.

Step 4. Repeat Steps 2 and 3 n_{sim} times, store the optimal weights and calculate expected utility according to

$$\begin{aligned}\text{EU}(\boldsymbol{w}_m^*) &= E(\boldsymbol{w}_m^{*\mathrm{T}}\boldsymbol{\mu}_0 - \tfrac{1}{2}\lambda\boldsymbol{w}_m^{*\mathrm{T}}\boldsymbol{\Omega}_0\boldsymbol{w}_m^*) \\ &= \frac{1}{n_{\text{sim}}}\sum_{j=1}^{n_{\text{sim}}}(\boldsymbol{w}_{m,j}^{*\mathrm{T}}\boldsymbol{\mu}_0 - \tfrac{1}{2}\lambda\boldsymbol{w}_{m,j}^{*\mathrm{T}}\boldsymbol{\Omega}_0\boldsymbol{w}_{m,j}^*)\end{aligned} \tag{7.9}$$

and compare all three methodologies with respect to Equation 7.9.

Let us summarise the above process. We start with an assumed knowledge of the truth, $\boldsymbol{\mu}_0$. Next we offer all three algorithms a total of n_{sim} drawings ($\bar{\boldsymbol{\mu}}_j, j = 1, \dots, n_{\text{sim}}$) from the truth. For each drawing the algorithm constructs an optimal portfolio. While Markowitz optimisation will use $\bar{\boldsymbol{\mu}}_j$ directly, our Bayesian alternative will apply Equation 7.7 to formulate a guess about $\boldsymbol{\mu}_0$. Michaud optimisation instead will repeatedly draw another $i = 1, \dots, k$ sample for each $\bar{\boldsymbol{\mu}}_j$ to derive optimal allocations. All n_{sim} allocations (for each of the three methodologies) will then be evaluated using Equation 7.9.

The above procedure is applied with one exemption to the original data set of Michaud (1998, Tables 2.3 and 2.4). The exemption is that we only use the equity data, ie, we restrict the universe to six rather than eight assets. The reason for this is twofold. First, shrinkage to a common return level, is only plausible within a homogeneous group of assets. Second, while we could have assumed a data set, the results are much more convincing, if we use Michaud's own data and it also allows an easier comparison with Markowitz and Usmen (2003). Finally, given the return dispersion in Michaud's data (returns range from 39bp to 88bp a month) an equal return prior does not look like an in-sample optimisation. We have made our lives more difficult, rather than easier.

7.4 OUT-OF-SAMPLE RESULTS: UNCONSTRAINED OPTIMISATION

We start with the case of unconstrained optimisation, ie, investors can go long and short any of the $n = 6$ assets without constraints. This case is interesting for two reasons. First long–short investing applies to many hedge funds or overlay managers that are virtually unrestricted in weights space. That does not mean they can put on unlimited leverage as these investment strategies are run under (active) risk constraints. Measuring risk directly (tracking risk, variance, etc) is in any case more flexible and precise than indirect measures like leverage. Second, we know that the impact of estimation error (not estimation error itself) on optimal portfolios is largest for unconstrained portfolios. It is this situation, where investors most need a safeguard.

The results for unconstrained portfolio construction are summarised in Table 7.1 for the original data set in Michaud (1998).

Table 7.1 Expected utility and risk aversion (w unconstrained)

	λ_1 = 0.02	λ_2 = 0.05	λ_3 = 0.1
$E(U_{\text{st}} - U_{\text{res}})$	1.54 (25.01)	3.85 (25.01)	7.7 (25.01)
$E(U_{\text{st}} - U_{\text{mv}})$	1.54 (25.02)	3.85 (25.02)	7.8 (25.02)
$E(U_{\text{res}} - U_{\text{mv}})$	−0.00 (−0.29)	−0.00 (−0.29)	−0.01 (−0.29)

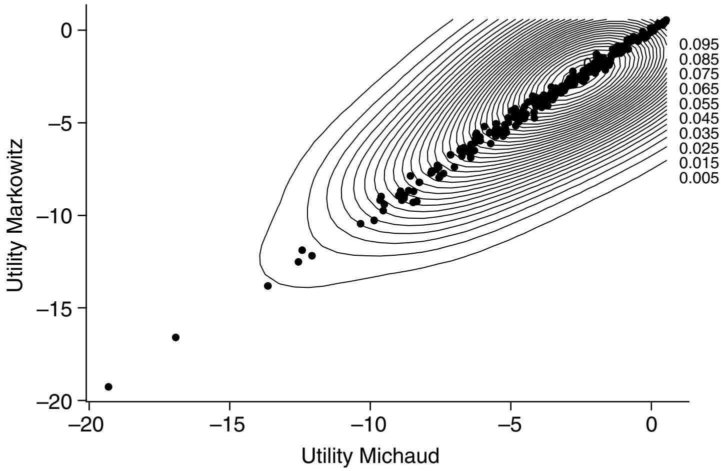
We compare all three portfolio construction techniques across three levels of aggressiveness. The first number shows the difference in expected utility which has (for the assumed utility function) a return dimension. A value of 1.54 means a monthly advantage of 154bp (due to leverage). The second value (in parentheses) is the t value of this difference. High t values (above 2) indicate a significant difference (at the 5% level).

We have chosen $\tau = 60$, $k = 250$ and $n_{\text{sim}} = 250$. Risk aversions can take three values: 0.02, 0.05, 0.1.[13] All optimisations are run unconstrained, ie, are solved using simple matrix algebra as in Equation 7.3. First we notice, what we already suspected from our analytical treatment. The difference between Markowitz and RE is negligible (t value of −0.29). As a consequence both methods lead to portfolios that only differ in noise as can be seen in Figure 7.1. Confirmation for this can also be found in Figure 7.1.

The second conclusion from the above out-of-sample test is that Bayesian return shrinkage is vastly superior to Markowitz and Michaud optimisation in the absence of long-only constraints. The results are both economically meaningful as well as statistically significant. James–Stein estimation outperforms RE by 154bp per month (as measured by the difference in expected utility) and this difference is not due to chance (t value of above 25).[14] We could stop with the above experiment. However, it may be argued that the results critically depend on the particular choice of $\boldsymbol{\mu}_0$. We agree. That is what makes out-of-sample tests of a particular set of Bayesian priors impossible to generalise. In other words we have aimed to show how easy it is to reverse the results of Markowitz and Usmen (1983) even within the same data set and investment universe.

One way to weaken (but not solve) this problem is to add a series of new experiments by drawing a new "truth", $\tilde{\boldsymbol{\mu}}_0$, from $\boldsymbol{\mu}_0$. But since

Figure 7.1 Markowitz versus resampling



All out-of-sample realisations of utility scatter around an (imaginary) straight line. Deviations are purely random.

Table 7.2 Additional out-of-sample experiments (unconstrained investor)

Experiment	$E(U_{\text{st}} - U_{\text{res}})$	t value
# 1	2.80	7.04
# 2	1.55	4.23
# 3	0.87	1.53
# 4	0.94	2.64
# 5	1.23	2.65
# 6	1.55	3.26
# 7	1.59	5.10
# 8	1.62	6.52
# 9	0.82	2.07
# 10	1.36	3.18

All calculations have been performed for an investor with $\lambda = 0.02$. Utility differences scale with risk aversion. For $\lambda = 0.1$ the difference is only one fifth of the above difference. However, t values remain unchanged.

all samplings from $\tilde{\boldsymbol{\mu}}_0 \sim N(\boldsymbol{\mu}_0, \upsilon^{-1}\boldsymbol{\Omega}_0)$ will be inevitably related to $\boldsymbol{\mu}_0$ (and hence inherit some of its arbitrariness) we can only ever partially address the problem. The results for 10 further experiments (with $\upsilon = 216$) are summarised in Table 7.2.

Again the outperformance of Bayes-based portfolio construction is overwhelming. Our results lead us to the conclusion that RE does badly relative to our Bayesian alternative, when we need it most, ie, when we have the ability to leverage on tiny estimation errors by building seemingly riskless long–short hedge portfolios. Note, also that our results have been independent from the assumed true distribution. In any of the ten subexperiments we confirm the inferiority of resampling versus shrinkage. This will come as no surprise to readers familiar to Bayesian methods. James–Stein estimators applied to Markowitz optimisation tend consistently to do better than maximum likelihood estimators combined with Markowitz optimisation.[15]

7.5 OUT-OF-SAMPLE RESULTS: CONSTRAINED OPTIMISATION

Unfortunately, the analytic simplicity of the previous section does not carry over to the case of constrained portfolio optimisation, where no closed-form solutions are available. We can no further equalise Markowitz portfolios and RE, nor can we expect Bayesian shrinkage to work so well. Even after applying shrinkage, some estimation errors will remain in the data which will in turn lead to some degree of error maximisation.

We start again with the original data set in Michaud (1998). Rather than in the previous section we now use a long-only constraint ($\boldsymbol{w} \geqslant \boldsymbol{0}$) as well as an adding-up constraint ($\boldsymbol{w}^{\mathrm{T}}\mathbf{1}_n = 1$). The results are summarised in Table 7.3. For a start we realise that out-of-sample performance differences between all three methodologies are much smaller than before. This is entirely consistent with the literature, where it is well known that long-only constraints (position limits) work as a safeguard against estimation error because they do not allow "almost riskless" long–short positions (that are most likely) the results of estimation error.

Second, we see that again shrinkage has been superior to both resampling as well as naive Markowitz. Note that this happens against the background of an unfavourable prior given our knowledge of the truth.

The difference between James–Stein shrinkage and resampling, however, becomes smaller with falling risk aversion. Portfolios built

Table 7.3 Expected utility and risk aversion (long-only constraint)

	$\lambda_1 = 0.02$	$\lambda_2 = 0.05$	$\lambda_3 = 0.1$
$E(U_{\mathrm{st}} - U_{\mathrm{res}})$	0.01 (2.94)	0.08 (17.18)	0.12 (20.36)
$E(U_{\mathrm{st}} - U_{\mathrm{mv}})$	0.07 (11.97)	0.14 (17.92)	0.14 (18.37)
$E(U_{\mathrm{res}} - U_{\mathrm{mv}})$	0.06 (16.71)	0.05 (12.66)	0.01 (5.71)

RE beats naive Markowitz, but again underperforms Markowitz optimisation combined with return shrinkage.

aggressively will still suffer from error maximisation as shrinkage cannot completely eliminate estimation error. The same applies to naive Markowitz. It does best (but still underperforms RE) for high risk aversions. We can summarise, that RE underperforms our Bayesian alternative. Again the result is opposite to Markowitz and Usmen (2003).

As before our results might be due to a deliberately engineered (but unrepresentative) choice of $\boldsymbol{\mu}_0$. We again attempt to soften this problem by adding a new series of experiments, drawing ten new "truths", $\tilde{\boldsymbol{\mu}}_0$, from $\boldsymbol{\mu}_0$. The results are summarised in Tables 7.4–7.6. The additional experiments confirm our previous results. Bayesian adjustments outperform resampling in 26 out of 30 cases. However, out-of-sample performance seems also to depend on risk aversion. If risk aversion is high, resampling is dominated by our Bayesian adjustment. This result is both statistically as well as economically significant. For lower risk aversions we find that while resampling underperforms on average, the difference is economically not very significant (2bp a month on average).

7.6 SUMMARY

The concept of RE yields underperforms a Bayesian alternative dramatically if position limits are absent. This casts serious doubts on RE as it does not work, when we need it most, ie, when we have the ability to leverage on tiny estimation errors by building seemingly riskless long–short hedge portfolios. Our results have been

Table 7.4 Additional experiments for aggressive investor ($\lambda = 0.02$)

#	$E(U_{\text{st}} - U_{\text{res}})$	t	$E(U_{\text{st}} - U_{\text{mv}})$	t	$E(U_{\text{res}} - U_{\text{mv}})$	t
1	0.08	14.00	0.13	14.00	0.05	7.90
2	0.05	11.00	0.01	1.40	0.06	11.00
3	−**0.06**	−**11.00**	−**0.04**	−**4.10**	**0.02**	**3.00**
4	0.05	8.50	0.05	6.00	0.01	1.10
5	0.12	20.00	0.15	14.00	0.03	3.60
6	0.03	6.30	0.05	7.20	0.02	3.90
7	0.08	11.00	0.10	8.30	0.02	2.10
8	−**0.05**	−**9.70**	−**0.01**	−**0.71**	**0.04**	**7.00**
9	−**0.09**	−**15.00**	−**0.07**	−**7.40**	**0.02**	**4.10**
10	0.08	12.00	0.09	7.40	0.01	0.94

'#' denotes the experiment number; 't' denotes t value. James–Stein return estimates (explicit shrinkage) outperforms RE (implicit shrinkage) most of the time (7 out of 10 experiments). Resampling does very well against the Markowitz alternative (10 out of 10 experiments).

Table 7.5 Additional experiments for intermediate risk aversion ($\lambda = 0.05$)

#	$E(U_{\text{st}} - U_{\text{res}})$	t	$E(U_{\text{st}} - U_{\text{mv}})$	t	$E(U_{\text{res}} - U_{\text{mv}})$	t
1	0.13	20.00	0.16	16.00	0.03	5.50
2	0.04	9.90	0.11	15.00	0.07	16.00
3	0.00	0.14	0.05	6.80	0.05	12.00
4	0.08	14.00	0.08	8.70	−0.01	−1.10
5	0.15	21.00	0.17	14.00	0.02	2.60
6	0.07	13.00	0.08	9.80	0.01	2.80
7	0.12	17.00	0.12	11.00	−**0.01**	−**0.96**
8	0.02	3.80	0.08	11.00	0.06	15.00
9	−**0.02**	−**3.80**	0.04	6.00	0.06	15.00
10	0.11	14.00	0.11	9.50	0.00	0.51

'#' denotes the experiment number; 't' denotes t value. James–Stein return estimates (explicit shrinkage) outperforms RE (implicit shrinkage) in 9 out of 10 experiments. Resampling does also well against the Markowitz alternative (8 out of 10 experiments).

independent from the assumed true distribution. In all of our subexperiments we confirm the inferiority of resampling versus explicit shrinkage.

Given the poor performance of resampling in an unconstrained setting and given the fact that long-only constraints provide some

Table 7.6 Additional experiments for risk-averse investor ($\lambda = 0.1$)

#	$E(U_{st} - U_{res})$	t	$E(U_{st} - U_{mv})$	t	$E(U_{res} - U_{mv})$	t
1	0.13	20.02	0.13	17.57	0.00	−0.07
2	0.10	19.17	0.12	18.02	0.02	7.03
3	0.06	11.34	0.08	13.10	0.03	9.87
4	0.07	12.65	0.06	9.88	−**0.01**	−**5.30**
5	0.13	17.89	0.12	15.32	−**0.01**	−**3.61**
6	0.07	12.37	0.07	9.56	−**0.00**	−**1.60**
7	0.10	14.96	0.07	9.53	−**0.03**	−**9.03**
8	0.07	13.88	0.10	14.73	0.02	7.49
9	0.05	9.95	0.07	10.72	0.02	7.94
10	0.09	12.81	0.08	9.94	−**0.01**	−**4.26**

‘#’ denotes the experiment number; ‘*t*’ denotes *t* value. James–Stein return estimates always outperform RE (ten out of ten experiments). Resampling offers no advantage relative to Markowitz optimisation as we move closer to the minimum variance portfolio (which has been known with certainty in the current setup).

sort of shrinkage, this in turn means that whatever resampling achieves (well established outperformance versus naive Markowitz optimisation) must be closely related to the positive effect of position limits on out-of-sample performance. When we investigate this case we have been able to reverse the results by Markowitz and Usmen (2003). In our simulation set up, resampling underperforms relative to Bayesian methods. RE continues, however, to outperform naive Markowitz, even though the exact mechanics remain unclear (and unformulated). It did not need much “engineering” from the authors’ side to arrive at these results, as we essentially used Michaud’s own set of data. This experiment proves a more general point. Out-of-sample testing of Bayesian methods is conceptually difficult. For every distribution we will find a prior that will outperform resampling (and vice versa), although we belief that a shrinkage estimator is a natural choice.

EXERCISE

1. Replicate Table 7.3 by downloading the free Excel addon poptools.xla and run a Monte Carlo simulation on your spreadsheet.

[1] An excellent review can be found in Connor *et al* (2010).

[2] Note that Bayesian analysis does not allow any additional adjustments ("hand waving") after parameter uncertainty has been addressed using priors. Resampling the predictive distribution is inconsistent with Bayesian theory. In other words, we cannot bring together Bayesian methods and RE.

[3] US patent #6003018.

[4] Jorion (1992) already used resampling to visualise the effect of estimation error on the distribution of portfolio weights. Simply averaging across this weights for a given risk aversion will lead to RE.

[5] For a comprehensive review of resampling see Scherer (2002).

[6] Even though Markowitz optimisation has been described so many times by authors other than him it is still illuminating to read Markowitz (1959).

[7] The use of averaging maintains portfolio properties, ie, averaged weights still add up to 100% (the same does not apply to median weights). At the same time averaging induces a bias into optimal portfolios as the weight distribution might become very asymmetric in the presence of constraints in which case the average is a poor location parameter.

[8] Note that we need to average across portfolios that have been constructed using an investors risk aversion, rather than using rank based averaging (average across portfolio that carry the same rank, ie, is the 20th portfolio out of a fixed number of portfolios ranging between the minimum risk and maximum return portfolio) suggested by Michaud. This is the only way to make consistent choices across environments with differing risk return characteristics. To put it differently: Portfolio #30 will carry a different risk/return trade-off in a high versus low risk premium environment.

[9] In contrast to Bayesian analysis this limits us to provide an assessment of the whole data set, rather than a subset. To put it differently: we cannot express more uncertainty about hedge fund returns, rather than government bonds if statistics are calculated from the same sample size. This makes sense only if we assume that all uncertainty arises from having too little data.

[10] Essentially this means that unconstrained investors (like hedge funds) might want to restrain from resampling, as it provides them with a slow and noisy version of naive Markowitz optimisation.

[11] Ceria and Stubbs (2005).

[12] Shrinking returns to whatever constant will not alter the maximum information ratio portfolio as the relative returns remain unchanged. However, investors will choose a less risky portfolio as for a given level of risk aversion the inclusion of estimation risk made the world a riskier place. The new optimal portfolio will lie between the maximum information ratio and the minimum risk portfolio.

[13] We do not express returns as percentages, ie, the implicit risk aversion for a portfolio with mean 10% and volatility 15% becomes $(10/225) = 0.044$.

[14] We also see that the difference in expected utility for $\lambda_1 = 0.02$ is essentially five times larger than for $\lambda_2 = 0.1$, which is a direct effect from leverage. In other words the optimal weights differ by a factor of $\lambda_2/\lambda_1 = 5$.

[15] See Kempf and Memmel (2003) for a formal proof.

REFERENCES

Ceria, S., and R. Stubbs, 2005, "Incorporating Estimation Error into Portfolio Selection: Robust Efficient Frontiers", Axioma Working Paper.

Connor, G., L. R. Goldberg and R. A. Korajceck, 2010, *Portfolio Risk Forecasting* (Princeton, NJ: Princeton University Press).

Jagannathan, R., and T. Ma, 2003, "Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps", *Journal of Finance* 43(4), pp. 1651–83.

James, W., and C. Stein, 1961, "Estimation with Quadratic Loss", in *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, vol. 1, pp. 361–79 (University of California Press).

Jorion, P., 1992, "Portfolio Optimization in Practice", *Financial Analysts Journal* 48, pp. 68–74.

Kempf, A., and C. Memmel, 2003, "Parameterschätzungen in der Portfoliotheorie: Ein analytischer und simulationsgestützter Vergleich", *Die Betriebswirtschaft* 63, pp. 516–31.

Markowitz, H., 1959, *Portfolio Selection: Efficient Diversification of Investments* (New York: John Wiley and Sons; 1991, 2nd edn, Cambridge, MA: Basil Blackwell).

Markowitz, H., and N. Usmen, 2003, "Diffuse Priors vs. Resampled Frontiers: An Experiment", *Journal of Investment Management* 1(4), pp. 9–25.

Michaud, R. O., 1998, *Efficient Asset Management: A Practical Guide to Stock Portfolio Optimization and Asset Allocation* (New York: Oxford University Press).

Scherer, B., 2002, "Portfolio Resampling: Review and Critique", *Financial Analysts Journal* 45(6), pp. 98–108.

8

Portfolio Construction with Transaction Costs

8.1 TRANSACTION COSTS

So far we have assumed that all positions within a given portfolio can be created from cash at zero transaction cost. This assumption will be relaxed in this chapter, where we focus on the costs associated with buying and selling securities. Although we use standard mean–variance optimisation to describe investors' preferences, the analysis can be extended to any other objective function treated in this book. It is well known that the average active investor will, by definition, exhibit zero alpha before costs and underperform after costs.[1] Hence, the inclusion of transaction costs in the portfolio construction process is material and has been attracting increasing attention.

Transaction costs come in many forms. In general, we can identify three sources.

Brokerage commission. This reflects the mechanics of order processing and is an explicit fee charged by the broker to handle the trade. Commissions have been under constant pressure and reflect only a very small part of transaction costs.

Bid–ask spread. The bid–ask spread reflects the costs of buying a security and selling it immediately after. Economically, we expect the bid–ask spread to rise with the costs of inventory (for the trader who is providing liquidity) as well as with asymmetrical information (the counterparty initiating a trade might have insider information) or uncertainty (the market might move against the liquidity provider before they are able to unload their position).

Market impact. The bid–ask spread will change depending on the volume traded. If, for example, the volume is large relative to the average daily volume, it will take longer for the liquidity provider

> to unload their position without impacting the market. The bid–ask spread will therefore rise. For large asset managers with scalable investment processes that create all accounts to trade in the same direction market impact is the most important source of transaction costs.

The literature often suggests that transaction cost, tc, is of the following functional form

$$\text{tc} = \text{Commission} + \frac{\text{Bid}}{\text{Ask}} - \text{Spread} + \theta\sqrt{\frac{\text{Trade volume}}{\text{Daily volume}}} \tag{8.1}$$

where Bid/Ask is expressed as a percentage and θ is a constant that needs to be estimated.[2] Though very intuitive, we often cannot use it for all assets within a portfolio as we need data on daily trade volumes. For currency forwards, for example, this number simply does not exist.

Alternatively we could see transaction costs as opportunity costs. How much performance does an investor lose if they do not trade immediately, but rather trade over time to eliminate market impact? We could model this as an average price option (Asian), where the option value would rise with volatility as well as with time horizon. Larger blocks would need longer time periods to be traded without impacting the market. The investor will purchase any given volume at an average price.

However, we will follow neither of these approaches. Instead, it will be assumed that the functional form of transaction costs is already given, ie, we are not concerned with the econometrics of estimating transaction costs. We rely on the practical assumption that real transaction costs can be approximated using piecewise linear functions that have already been provided by the respective counterparties. This is very often the case, as we see from Table 8.1. Transaction costs rise with the number of futures contracts traded. Small markets (IBEX 35) require substantial transaction cost adjustments relative to large markets (DAX).

8.2 TURNOVER CONSTRAINTS

Turnover constraints are implemented by practitioners to provide a heuristic safeguard against transaction costs. The implicit assumption behind this indirect treatment of transaction costs is that, if transaction costs are proportional and equal across assets, it is sufficient

Table 8.1 Transaction cost estimates (%)

	Number of contracts traded						
	< 20	**20–50**	**50–100**	**100–250**	**250–500**	**1,000**	**2,000**
AEX 25 Index futures	0.00	0.03	0.08	0.25	0.40	0.50	0.80
IBEX 35 Index futures	0.12	0.25	0.30	0.43	0.75	—	—
DAX futures (EUX)	0.02	0.02	0.02	0.06	0.10	0.20	0.30
DJ Euro Stoxx 50 (EUX)	0.03	0.03	0.03	0.03	0.03	0.03	0.07
FTSE 100 futures (LIFFE)	0.02	0.02	0.03	0.05	0.09	0.16	0.26
MIB 30 futures (MSE)	0.01	0.01	0.03	0.09	0.18	0.40	0.75
SMI futures (EUZ)	0.01	0.01	0.03	0.06	0.08	0.14	0.21
CAC 40 Index futures	0.02	0.04	0.07	0.07	0.14	0.20	0.30
OMX futures (OML)	0.04	0.04	0.08	0.06	0.15	0.25	0.40
South Africa All Share (SAF)	0.00	0.05	0.10	0.20	0.50	1.00	1.50

to control turnover as it relates directly to transaction costs. In reality, transaction costs differ across assets and do not change proportionally with trade size. Nevertheless, we will use turnover constraints as a starting point.

So far it has not been necessary for us to know the initial holdings, w_i^{Initial}, when constructing a portfolio as we assumed that no costs were involved in turning our portfolio into cash (and vice versa). However, this needs to change now. In addition to the vector of initial holdings we need:

1. two new sets of variables, ie, assets bought, w_i^{+} (positive weight changes), and assets sold, w_i^{-} (negative weight changes);
2. a budget constraint for each asset that requires that the final asset weight, w_i, equals the initial weight, w_i^{Initial}, plus transactions: $w_i = w_i^{\text{Initial}} + w_i^{+} - w_i^{-}$; and

3. the turnover constraint itself, which limits the total turnover, τ, that is permitted in a portfolio to a specified number

$$\sum_{i=1}^{n}(w_i^+ - w_i^-) \leqslant \tau, \quad w_i^+ \geqslant 0,\ w_i^- \geqslant 0$$

The corresponding portfolio optimisation problem can be expressed as

$$\begin{aligned}
&\min \sum_i \sum_j w_i w_j \sigma_{ij} \quad \text{subject to}\\
&\sum_i w_i \mu_i = \mu\\
&\sum_i w_i = 1\\
&w_i = w_i^{\text{Initial}} + w_i^+ - w_i^-\\
&\sum_{i=1}^{n}(w_i^+ - w_i^-) \leqslant \tau\\
&w_i^+ \geqslant 0\\
&w_i^- \geqslant 0\\
&w_i \geqslant 0
\end{aligned} \tag{8.2}$$

System 8.2 can now be implemented in the portfolio optimisation software of your choice.[3] Turnover constraints effectively anchor the optimised portfolio around the initial holdings. Small turnover figures allow only small deviations, while the reverse is true for large turnover figures.

Although turnover figures attract a lot of attention in account review meetings, this emphasis is largely unjustified for various reasons.

- Turnover constraints do not allow an optimal treatment of transaction costs as these may vary considerably across asset classes. High turnover in fixed-income futures creates very little transaction cost and is nothing to worry about from a cost perspective, while this is not the case when trading small caps.
- It does not always make sense to enforce turnover limits. If an asset manager has already spent their turnover budget of 40% per annum by June, should they stop trading even though there are still investment opportunities around? An investor with

forecasting skills who acts on a signal that arrives quite often with a quick information decay might want high turnover. Turnover limits reduce the amount of risk taken, but as long as turnover is evenly distributed around the trading horizon the information ratio remains undamaged.

- High turnover might create neither excessive risks nor large transaction costs. Long–short positions in highly correlated index futures are cheap to implement (and require a large size until they impact an investment manager's risk budget).
- Constraining turnover by trading only on large signals (see the Michaud resampling approach discussed in Chapter 4) violates the fundamental law of active management by destroying diversification. The breadth of active positions is seriously limited if only a fraction (the most extreme) of positions is implemented.

8.3 TRADING CONSTRAINTS

Another way to deal with turnover is to place constraints on the trading of individual assets. Often the idea is to trade only those positions that are deemed to be economically meaningful. The most prominent examples are as follows.

Lower bound on weights. Portfolio managers (and their clients) often hate small, active positions (deviations from benchmark holdings) which, they argue, contribute little to total performance. Hence, we could enforce this view by adding the constraint that, if an active position is established, it must be at least x% of total volume. This is by its very nature a go/no-go decision.

Upper bound on number of assets. Portfolio diversification is helpful, but too much diversification might increase transaction costs as well as monitoring costs. We could therefore limit the number of assets to a specified maximum so as to ensure a manageable portfolio. Assets are counted as 0 or 1 depending on whether they are in the solution set or not.

The modelling of these problems is not very widespread. Little commercial software is available as it involves integer constraints in combination with a non-linear objective function.

Table 8.2 Formulation of buy-in thresholds

Type	Formula
Either in or out	$w_i \leqslant \delta_i \cdot$ Large number, $\delta_i = 0, 1$
Either in between or out	$\delta_i w_i^{\min} \leqslant w_i \leqslant \delta_i \cdot w_i^{\max}$, $\delta_i = 0, 1$
Either out or above	$w_i \leqslant \delta_i \cdot$ Large number, $w_i^{\max}\delta_i \leqslant w_i$, $\delta_i = 0, 1$
Cardinality constraint	$\sum_i \delta_i =$ Number of assets

Buy-in thresholds are an example of maximum bounds on the number of positions. Suppose we need to determine optimal portfolio weights, w_i. However, we also want to restrict the number of assets we invest into. Note that a 0.0001% weight will contribute the same to the count of assets as a 10% weight. If we introduce a new binary variable, δ_i, that assumes 1 if an asset is included in the optimal solution and 0 otherwise, we are able to model the inclusion/exclusion of individual assets

$$\delta_i = \begin{cases} 1 & \text{if asset } i \text{ is selected} \\ 0 & \text{otherwise} \end{cases} \tag{8.3}$$

Remember we want to count how many assets enter the optimal solution. As an asset is either in or out of the optimal solution, we can express this using

$$w_i \leqslant \delta_i \cdot \text{Large number}, \quad \delta_i = 0, 1 \tag{8.4}$$

Equation 8.4 operates like a switch. If an asset is included (even in a very small amount) the inequality is only satisfied for $\delta_i = 1$. As soon as the asset leaves the solution set, Equation 8.4 holds only for $\delta_i = 0$. Computationally, what is referred to here as a “large” number should not be made too large. We can extend this logic to model typical buy-in thresholds as summarised in Table 8.2.

It is well known that diversifying into a broader universe of assets has both merits and limitations. Positions may become very small, monitoring and research costs rise and diversification benefits tend to fade away as the number of assets becomes large. Thus, investors might want to solve the following mixed-integer quadratic program

$$\min \sum_i \sum_j w_i w_j \sigma_{ij} \quad \text{subject to}$$

$$\sum_i w_i \mu_i = \mu$$

$$\begin{aligned}
&\sum_i w_i = 1 \\
&\delta_i w_i^{\min} \leqslant w_i \leqslant \delta_j \cdot w_i^{\max} \\
&\sum_i \delta_i = \text{Number of assets} \\
&w_i \geqslant 0 \\
&\delta_i \in \{0, 1\} \qquad (8.5)
\end{aligned}$$

We apply 8.5 to a set of sample data drawn from a normal distribution (50 draws) with means 2%, 4%, 5% and 8%. All assets exhibit 20% volatility (with zero correlation). There are $s = 1, \dots, 50$ scenarios for portfolio returns, r_{ps}, that are constructed from individual asset returns, r_{is}

$$\sigma_{\mathrm{p}} = \sqrt{\frac{1}{n} \sum_s (r_{\mathrm{ps}} - \bar{r}_p)} \quad \text{and} \quad r_{\mathrm{ps}} = \sum_{i=1}^{4} w_i r_{\mathrm{is}} \qquad (8.6)$$

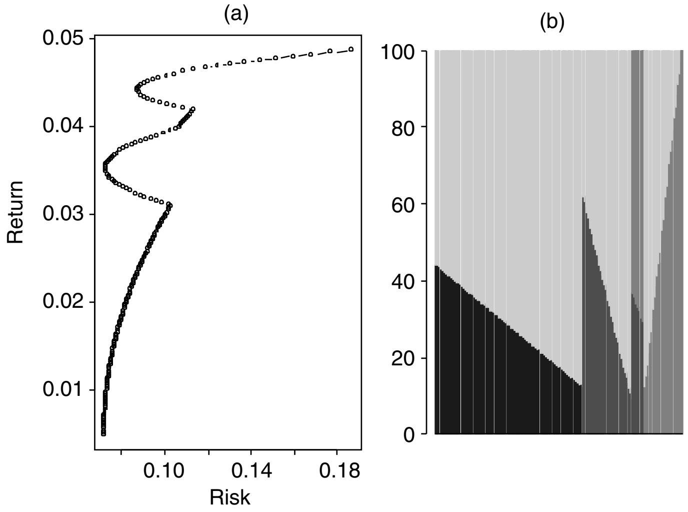
At first sight the result is an odd-looking frontier. Figure 8.1 is best understood if we think of the frontier as the envelope for all possible pairwise combinations (cardinality constraint of 2) within the four-asset universe (six combinations). Each shade in the figure represents a different asset. Note that tracing out the frontier by increasing the return requirements stepwise will lead to inefficient parts. In reality the frontier becomes discontinuous as it makes little sense to invest into dominated portfolios (which are exposed to higher risk for the same expected return). Needless to say that the cardinality-constrained portfolio plots below an efficient frontier without cardinality constraints. The difference tends to be largest at the minimum-variance portfolio (where diversification normally requires many assets to be included as long as the assets have similar risk characteristics), and it tends to be zero at the maximum-return portfolio, which tends to be concentrated in a single asset (the maximum-return asset).

We can also use the above logic to model round lots (where stocks can only be purchased in blocks). As before, we model the holdings in the ith asset as

$$w_i = \delta_i \, \mathrm{Block}_i$$

where Block_i = Round lot transaction for asset_i/Total wealth. Note, however, that the number of blocks, δ_i, times the block sizes, Block_i,

Figure 8.1 Efficient frontier for cardinality-constrained portfolios



does not need to sum to one. We can accommodate this by introducing over- and undershoot variables, ϖ^+, ω^-, into the budget constraint

$$\sum_{i=1}^{n} \delta_i \, \mathrm{Block}_i + \varpi^+ - \varpi^- = 1 \tag{8.7}$$

Over- and undershoots need to be penalised in the objective function with an "appropriate" cost factor. Our portfolio optimisation model now becomes

$$\begin{aligned}
&\min \sum_i \sum_j w_i w_j \sigma_{ij} + c(\varpi^+ + \varpi^-) \quad \text{subject to} \\
&\sum_i w_i \mu_i = \mu \\
&w_i = \delta_i \, \mathrm{Block}_i \\
&\sum_i w_i + \varpi^+ - \varpi^- = 1 \\
&w_i \geqslant 0 \\
&\delta_i \in \{0, 1\} \\
&\varpi^+, \varpi^- \geqslant 0
\end{aligned} \tag{8.8}$$

The cost factor c needs to be chosen carefully.

A related problem is to track a given index with a small number of stocks. Trading desks at investment banks very often face a similar problem. They are asked by clients to construct tracking baskets that are a group of stocks that is constrained in size to a maximum number of stocks. Let us denote the holdings of the target portfolio (benchmark portfolio to be tracked with minimum tracking error) by b_i. The problem of tracking an index with a small number of stocks now becomes

$$\begin{aligned}
\min \sum_i \sum_j (w_i - b_i)(w_j - b_j)\sigma_{ij} \quad & \text{subject to} \\
\sum_i w_i &= 1 \\
w_i &\leqslant \delta_i \cdot \text{Large number} \\
\sum_i \delta_i &= \text{Number of assets} \\
w_i &\geqslant 0 \\
\delta_i &\in \{0, 1\}
\end{aligned} \tag{8.9}$$

Again, this can be solved with some commercial optimisation packages (eg, NuOPT for S-Plus, Axioma, CPLEX).

8.4 PROPORTIONAL TRANSACTION COSTS

So far it has been cost-free to shift portfolio allocations.[4] However, there are transaction costs associated with buying and selling securities. We assume that transaction costs are paid at the beginning of the period. In order to model transaction costs we have to modify the budget constraint as the costs associated with our transactions have to be paid out of the existing budget, ie, they have to be financed from asset sales

$$\sum_{i=1}^{n}(w_i^- - w_i^+) - \sum_{i=1}^{n}(\mathrm{tc}_i^+ w_i^+ + \mathrm{tc}_i^- w_i^-) \geqslant 0, \quad w_i^+ \geqslant 0,\ w_i^- \geqslant 0 \tag{8.10}$$

where $\mathrm{tc}_i^{\pm}$ is the proportional transaction costs for buys and sells. Note that the first summation denotes the proceeds from net selling, while the second summation denotes the associated costs. Transaction costs lead to an indirect reduction of return as the amount on which asset returns can be earned is reduced from the start of the investment period, ie, $\sum_i w_i = \sum_i w_i^{\text{Initial}}$. We can incorporate Expression 8.10 into a new budget constraint by adding transaction costs to the summation of holdings that are left after transactions costs have been paid

$$
\begin{aligned}
&\min \sum_i \sum_j w_i w_j \sigma_{ij} \quad \text{subject to} \\
&\sum_i w_i(1+\mu_i) = 1 \\
&\sum_{i=1}^{n} w_i + \sum_{i=1}^{n} (\mathrm{tc}_i^+ w_i^+ + \mathrm{tc}_i^- w_i^-) = 1 \\
&w_i = w_i^{\text{Initial}} + w_i^+ - w_i^- \\
&w_i \geqslant 0 \\
&w_i^+ \geqslant 0 \\
&w_i^- \geqslant 0
\end{aligned}
\tag{8.11}
$$

Note that as now $\sum_i w_i \leqslant 1$ we need to write $\sum_i w_i(1+\mu_i) = 1+\mu$ instead of $\sum_i w_i\mu_i = \mu$. Since 8.11 is very similar to 8.2 we need only some minor changes to accommodate for transaction costs. Two interesting points are worth mentioning.

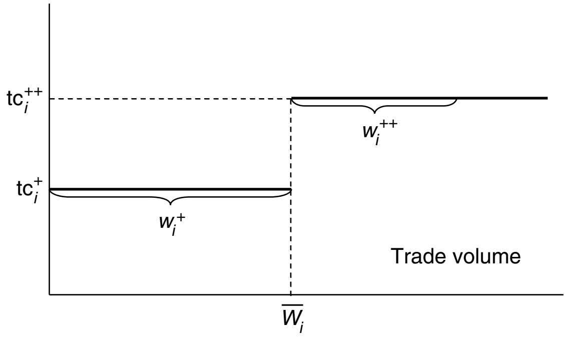
- Asset weights no longer add up to 100% (now 98.7%) as transaction costs are eating into initial wealth.
- Adjustments are made via either buys or sells, but no asset is bought and sold at the same time. Although this would have no direct effect on the net change in holdings, it would, however, induce transaction costs that were too high.

8.5 PIECEWISE LINEAR TRANSACTION COSTS

So far we have employed linear transaction costs, where costs per trade, tc_i^+ and tc_i^-, for the ith asset have been constant, ie, independent of trade size w_i^+, w_i^-. In reality, however, the trading costs per unit traded will rise with the volume traded (market impact), as we saw in Table 8.1.

It is assumed that we trade for an account of size W, where W denotes the portfolio value in millions. We will now modify our transaction cost function by assuming that trading costs rise as soon as we trade more than a critical volume $\bar{W}_i$, as illustrated in Figure 8.2.

Figure 8.2 Piecewise linear transaction costs



A straightforward change to 8.10 yields

$$\left.\begin{aligned}
&\sum_{i=1}^{n}(w_i^- + w_i^{--} - w_i^+ - w_i^{++})\\
&\qquad - \sum_{i=1}^{n}(\mathrm{tc}_i^+ w_i^+ + \mathrm{tc}_i^- w_i^-)\\
&\qquad\qquad - \sum_{i=1}^{n}(\mathrm{tc}_i^{++} w_i^{++} + \mathrm{tc}_i^{--} w_i^{--}) \geqslant 0\\
&w_i^+ \geqslant 0, \quad w_i^- \geqslant 0, \quad w_i^{++} \geqslant 0\\
&w_i^{--} \geqslant 0, \quad (w_i^+ - w_i^-)W \leqslant \bar{W}_i
\end{aligned}\right\} \tag{8.12}$$

We are now equipped to compare the optimisation results for a world without transaction costs as well as for linear and piecewise linear transaction costs. Suppose that initial positions are zero for all assets, and assume further that $\bar{W}_i$ is 30,000,000 for assets 1–3 and 5,000,000 for asset 4. We want to find a portfolio with a 4% expected return for an account with a volume of 100,000,000. Transaction costs are 1% (for buys and sells) in the case of linear transaction costs. For piecewise linear transaction costs the following table applies.

Asset	tc_i^+	tc_i^{++}
1	1%	2%
2	1%	2%
3	1%	2%
4	1%	4.5%

Table 8.3 Optimisation results

	Allocation					
Optimisation	**Asset 1**	**Asset 2**	**Asset 3**	**Asset 4**	$\sum_i w_i$	**Risk (%)**
No TC	0.33	0.39	0.20	0.08	1.000	2.4
Linear TC	0.02	0.54	0.32	0.11	0.991	3.4
Piecewise TC	0.00	0.33	0.54	0.12	0.985	4.3

TC, Transaction costs.

The results are summarised in Table 8.3. As we see, transaction costs will considerably alter the decomposition of the optimal portfolio.

All three portfolios share an expected return of 4%. However, while the most cost-effective minimum risk portfolio (no TC) has an annual risk of 2.4%, the inclusion of transaction costs forces the optimiser to allocate into higher-yielding assets in order to recoup transaction costs that eat up returns. As a result portfolio risk increases. If we had maximised return for a given portfolio risk, transaction costs would have reduced expected return instead. Increasing transaction costs for asset 4 to 9.5% (after the break) would result in an allocation that only used asset 4 up to the point where transaction costs became too high; ie, the new allocation would look like this: $\boldsymbol{w}' = [0 \quad 0.32 \quad 0.63 \quad 0.05]$.

8.6 FIXED TRANSACTION COSTS

Fixed transaction costs are another go/no-go situation, that is, fixed costs arise as soon as we trade in a particular asset (independent of trade size), while they are zero if no trade takes place. We can thus model the budget constraint for a combination of fixed and proportional costs according to

$$\underbrace{\sum_{i=1}^{n} w_i}_{\text{Holdings}} + \overbrace{\sum_{i=1}^{n} (\delta_i^+ + \delta_i^-) \cdot f_i}^{\text{Fixed}} + \underbrace{\sum_{i=1}^{n} (\mathrm{tc}_i^+ w_i^+ + \mathrm{tc}_i^- w_i^-)}_{\text{Proportional}} \tag{8.13}$$

Final holdings plus fixed and proportional transaction costs have to add up to the initial budget. Fixed costs are added up with the use of integer variables that take on a value of one if trading takes place and zero otherwise. We assume fixed costs to be the same for all assets. Extensions are trivial. The variables in 8.13 are defined as follows

$$\begin{aligned}
&w_i^+ \leqslant \delta_i^+ \cdot \text{Large number}\\
&w_i^- \leqslant \delta_i^- \cdot \text{Large number}\\
&\delta_i^\pm \in \{0, 1\}\\
&w_i^+ \geqslant 0\\
&w_i^- \geqslant 0
\end{aligned} \tag{8.14}$$

For ease of notation we use constant fixed costs, f_i, across assets. An extension of the notation is trivial. The portfolio optimisation problem now becomes

$$\begin{aligned}
\min \sum_i \sum_j w_i w_j \sigma_{ij} \quad & \text{subject to}\\
&\sum_i w_i(1+\mu_i) = 1+\mu\\
&\sum_{i=1}^n w_i + \sum_{i=1}^n (\delta_i^+ + \delta_i^-) \cdot f_i \sum_{i=1}^n (\text{tc}_i^+ w_i^+ + \text{tc}_i^- w_i^-) = 1\\
&w_i = w_i^{\text{Initial}} + w_i^+ - w_i^-\\
&w_i \geqslant 0\\
&w_i^+ \leqslant \delta_i^+ \cdot \text{Large number}\\
&w_i^- \leqslant \delta_i^- \cdot \text{Large number}\\
&\delta_i^\pm \in \{0, 1\}\\
&w_i^+ \geqslant 0
\end{aligned} \tag{8.15}$$

Note that the existence of fixed transaction costs will, all things being equal, lead to a larger focus on a small number of assets where transactions will be performed. Across time, a small number of large adjustments will be preferable to a large number of infinitesimal rebalancing trades.

8.7 MULTIPLE ACCOUNTS WITH PIECEWISE LINEAR TRANSACTION

Suppose that we employ a quantitative portfolio construction process on n accounts. Assume further that they all differ in either constraints, investment universes, size or benchmarks. Individual portfolios are optimised sequentially. The total volume traded will only be known after all accounts have been traded. However, total transaction costs are a function of total volume. If costs per trade remain constant, this does not pose a problem. But as soon as costs per trade are discontinuous (ie, they jump at higher volumes) we face two problems:

1. costs per trade are known only after we know the total trade volume; and
2. optimisations of individual portfolios will generally not allocate costs correctly.

Without loss of generality we set the number of accounts equal to two ($k = 1, 2$). The transaction cost (τ) function is assumed to be equal across assets and given by $\tau = \theta(\Delta w)^{\gamma}$, where Δw denotes the weight change in a given asset. Marginal transaction costs are given by

$$\frac{\partial \tau}{\partial \Delta w} = \theta\gamma(\Delta w)^{\gamma-1}$$

For example, setting $\gamma = 1$, transaction costs become linear with constant marginal costs

$$\frac{\partial \tau}{\partial \Delta w} = \theta$$

Average transaction costs are given by

$$\frac{\tau}{\Delta w} = \theta(\Delta w)^{\gamma-1}$$

The simultaneous optimisation of both accounts requires to maximise

$$\left(\sum_{i=1}^{n} w_{1i}\mu_i + \sum_{i=1}^{n} w_{2i}\mu_i\right) - \left(\sum_{i=1}^{n} \theta(\Delta w_{1i}^{+} + \Delta w_{2i}^{+})^{\gamma} + \sum_{i=1}^{n} \theta(\Delta w_{1i}^{-} + \Delta w_{2i}^{-})^{\gamma}\right) \quad (8.16)$$

subject to a series of restriction given in 8.17–8.19. The objective 8.16 includes expected returns net of transaction costs. Note again that we distinguish between buys and sells of a given asset. This arises for legal reasons as asset managers must not internally cross trades between client accounts. A sell of 10% in asset 5 from account #1 together with a 5% buy account #2 will not result into a 5% sell, but rather a 10% sell ($\Delta w_{15}^{-} = -10\%$) plus a 5% buy ($\Delta w_{25}^{+} = -5\%$).

Again we could specify different transaction costs for each asset as well as for buys $(+)$ and sells $(-)$ separately[5]

$$\left.\begin{array}{llll} w_{1i} = w_{1i}^{0} + \Delta w_{1i}^{+} - \Delta w_{1i}^{-}, & & w_{2i} = w_{2i}^{0} + \Delta w_{2i}^{+} - \Delta w_{2i}^{-} & \\ \Delta w_{i1}^{+} \geqslant 0, & \Delta w_{i2}^{+} \geqslant 0, & \Delta w_{i1}^{-} \geqslant 0, & \Delta w_{2i}^{-} \geqslant 0 \end{array}\right\} \tag{8.17}$$

Transactions (buys and sells) are differentiated with the use of a set of equality and inequality constraints in 8.17. The initial position in each asset is denoted by w_{ki}^{0}. However, to keep the example simple we set $w_{ki}^{0} = 0$ for all i and k in our numerical example. Implicitly we assume both accounts are of equal size as we restrict ourselves to weight changes. Account sizes can either be accommodated for by incorporating size directly into transaction cost, ie, multiplying weight changes by the size of the underlying account, or by appropriately scaling risk up or down. Risk is measured the usual way calculating portfolio variance

$$\sum_{i}^{n} \sum_{j}^{n} w_{1i} w_{1j} \sigma_{ij} = \bar{\sigma}_{1}^{2}, \qquad \sum_{i}^{n} \sum_{j}^{n} w_{2i} w_{2j} \sigma_{ij} = \bar{\sigma}_{2}^{2} \tag{8.18}$$

Whether portfolio variance measures active or total risk depends whether weights sum to zero or one in our budget constraint 8.19. Here we assume an active optimisation (risk is measured as squared tracking error). This requires that the sum of all asset weights from $i = 1, \dots, n$ equals zero for a given k

$$\sum_{i=1}^{n} w_{1i} = 0, \qquad \sum_{i=1}^{n} w_{2i} = 0 \tag{8.19}$$

Equations 8.16–8.19 establishes a particular parameterisation of the aggregate trading problem described in the previous sections. Note that the separate optimisation of an individual account (trading in isolation) is just a special case of the above specification. All we need to do is to set the tracking error of all other accounts to zero.[6]

In order to understand the mechanics of our parameterisation the aggregate trading problem, we present a numerical example. Assume 2 accounts with equal investment universe of 10 uncorrelated assets, where asset volatility ranges between 11% and 20% (in ten equidistant steps) and expected returns vary between 5.5% and 10% (also in equal steps).

Zero transaction costs. Without transaction costs the transaction cost term is dropped from the objective function and nothing glues both accounts together. The joint optimisation becomes essentially the same as separate individual optimisations. No preference is given to any of both accounts as can be seen in Table 8.4. We start with the requirement that both accounts target a 5% tracking error. The information ratio for both accounts are equal

$$\mathrm{IR}_k = \frac{\sum_{i=1}^{n} w_{ik}\mu_i}{\bar{\sigma}_k} = 0.2999 \quad \text{for } k = 1, 2 \tag{8.20}$$

as we would expect for identical accounts. The amount of turnover

$$\sum_{i=1}^{n} |\Delta w_{1i}| = \sum_{i=1}^{n} |\Delta w_{2i}| \tag{8.21}$$

is also the same. What happens if both accounts require different tracking errors (due to client aggressiveness)? We assume $\bar{\sigma}_1 = 3$, $\bar{\sigma}_2 = 5$. Again information ratios are identical as the information ratio does not depend on leverage, which is the only difference between both accounts (the relative weighting of assets remains the same). In fact the relative amount of turnover (equal to leverage in this example) is the same as the relative tracking error $0.5548/0.9246 = 3/5$.

If we introduce a constraint on the weighting of highest return asset (#10), the information ratio of the constraint account drops as we would expect (the constraint distorts the optimal weighting of assets), but it does so for the individual as well as the combined optimisation in exactly the same way. The constrained account also trades less than the unconstrained.

Linear transaction costs. In the case of linear transaction costs both accounts are linked via the transaction cost term. However, as linear transaction costs exhibit constant marginal costs, trades in one account will not alter the marginal costs in the other account. Again the results of a combined optimisation will be identical to the results of consecutive separate optimisations. It is straightforward to see that transaction costs for account k are given by

$$\tau_k = \sum_{i=1}^{n} \theta \Delta w_{ki} \tag{8.22}$$

and identical. This results in a lowered information ratio

$$\mathrm{IR}_k = \frac{\sum_{i=1}^{n} w_{ik}\mu_i - \tau_k}{\bar{\sigma}_k} = 0.2817 \quad \text{for } k = 1, 2 \tag{8.23}$$

Table 8.4 Simultaneous optimisation of two accounts with zero transaction costs ($\theta = 0$)

	Account		
	#1	#2	#1 + #2
IR_1	—	0.2999	0.2999
IR_2	0.2999	—	0.2999
$\sum_{i=1}^{n} \lvert\Delta w_{i1}\rvert$	—	0.9246	0.9246
$\sum_{i=1}^{n} \lvert\Delta w_{i2}\rvert$	0.9246	—	0.9246
τ_1	—	—	—
τ_2	—	—	—
τ^+	—	—	—
τ^-	—	—	—

	Account		
	#1	#2	#1 + #2
IR_1	0.2999	—	0.2999
IR_2	—	0.2999	0.2999
$\sum_{i=1}^{n} \lvert\Delta w_{i1}\rvert$	0.5548	—	0.5548
$\sum_{i=1}^{n} \lvert\Delta w_{i2}\rvert$	—	0.9246	0.9246
τ_1	—	—	—
τ_2	—	—	—
τ^+	—	—	—
τ^-	—	—	—

	Account		
	#1	#2	#1 + #2
IR_1	0.2634	—	0.2634
IR_2	—	0.2999	0.2999
$\sum_{i=1}^{n} \lvert\Delta w_{i1}\rvert$	0.8946	—	0.8946
$\sum_{i=1}^{n} \lvert\Delta w_{i2}\rvert$	—	0.9246	0.9246
τ_1	—	—	—
τ_2	—	—	—
τ^+	—	—	—
τ^-	—	—	—

Constraints: $\bar{\sigma}_1 = \bar{\sigma}_2 = 5$ (top); $\bar{\sigma}_1 = 3$, $\bar{\sigma}_2 = 5$ (middle); $\bar{\sigma}_1 = \bar{\sigma}_2 = 5$, $w_{1,10} \leqslant 0$ (bottom).

A second mean to measure of what is going on with (total) transaction costs is to split transaction costs into those associated with buys

Table 8.5 Simultaneous optimisation of two accounts with linear transaction costs $\theta = \frac{1}{10}$, $\gamma = 1$)

	Account		
	#1	#2	#1 + #2
IR_1	0.2817	—	0.2817
IR_2	—	0.2817	0.2817
$\sum_{i=1}^{n} \lvert\Delta w_{i1}\rvert$	0.9037	—	0.9037
$\sum_{i=1}^{n} \lvert\Delta w_{i2}\rvert$	—	0.9037	0.9037
τ_1	0.0904	—	0.0904
τ_2	—	0.0904	0.0904
τ^+	0.0452	0.0452	0.0904
τ^-	0.0452	0.0452	0.0904

	Account		
	#1	#2	#1 + #2
IR_1	0.2817	—	0.2817
IR_2	—	0.2817	0.2817
$\sum_{i=1}^{n} \lvert\Delta w_{i1}\rvert$	0.5422	—	0.5422
$\sum_{i=1}^{n} \lvert\Delta w_{i2}\rvert$	—	0.9037	0.9037
τ_1	0.0542	—	0.0542
τ_2	—	0.0904	0.0904
τ^+	0.0270	0.0452	0.0723
τ^-	0.0270	0.0452	0.0723

	Account		
	#1	#2	#1 + #2
IR_1	0.2457	—	0.2457
IR_2	—	0.2817	0.2817
$\sum_{i=1}^{n} \lvert\Delta w_{i1}\rvert$	0.8762	—	0.8762
$\sum_{i=1}^{n} \lvert\Delta w_{i2}\rvert$	—	0.9037	0.9037
τ_1	0.0876	—	0.0876
τ_2	—	0.0904	0.0904
τ^+	0.0438	0.0452	0.0890
τ^-	0.0438	0.0452	0.0890

Constraints: $\bar{\sigma}_1 = \bar{\sigma}_2 = 5$ (top); $\bar{\sigma}_1 = 3$, $\bar{\sigma}_2 = 5$ (middle); $\bar{\sigma}_1 = \bar{\sigma}_2 = 5$, $w_{i,10} \leqslant 0$ (bottom).

(τ^+) and sells (τ^-): $\tau = \tau^+ + \tau^-$. In the case of linear transaction costs these will obviously be symmetric, a property that will be lost under asymmetric costs. If for example many positive weights are financed by a few negative weights, these negative weights must be individually large, which means they carry larger transaction costs due to the nonlinearity of the transaction cost function.

Nonlinear costs. This is the most interesting case as performance of both accounts is now mutually dependent. Any trading in one account will affect the marginal transaction costs and therefore optimality in the other account(s). While individual optimisation leads to an information ratio of 0.2760 (again lower than 0.2817) the information ratio under combined optimisation drops further to 0.2529. We can easily identify transaction costs as the reason for this. While transaction costs for account k are 0.1178, if we assume each account trades independently they really are much higher if both accounts trade together (0.2277 per account). Note that transaction costs per account are calculated by allocating total transaction costs to individual accounts depending on how much each account has traded

$$\tau_1 = \tau_1^+ + \tau_1^- = \sum_{i=1}^{n} \Delta w_{1i}^+ \theta (w_{1i}^+ + w_{2i}^+)^{\gamma-1} + \sum_{i=1}^{n} \Delta w_{1i}^- \theta (w_{1i}^- + w_{2i}^-)^{\gamma-1} \tag{8.24}$$

Investigating the split of τ_1 into τ_1^+ and τ_1^- we see that transaction costs on for negative weights are much larger than for positive weights which indicates a few short positions relative to less concentrated long positions. The drop in information ratio in the constraint case is about 9.8% for account #1 (from 0.2388 to 0.2174). This is larger than for account 2 (8.6% from 0.2758 to 0.2540) because of the nonlinear transaction costs involved. While transaction costs are rising for both managers, due to the smaller number of positions the quadratic cost term affects manager 1 more than manager 2.

In the extreme manager #1 might be constrained to take all their positions in three assets. If other managers also use some capacity the constrained account is much more severely hit. Why does this solution qualify as optimal? For a start maximising 8.16 subject to 8.17–8.19 is essentially the same as maximising combined utility. All client objectives enter with the same weight independent of client size or tracking error requirements (see the third column in the middle Table 8.6, which gives the same information ratio to both accounts that are with the exception of tracking error otherwise identical). If combined utility is maximised we arrive at Pareto optimality. No

Table 8.6 Simultaneous optimisation of two accounts with quadratic transaction costs ($\theta = 1$, $\gamma = 2$)

	Account		
	#1	**#2**	**#1 + #2**
IR_1	0.2760	—	0.2529
IR_2	—	0.2760	0.2529
$\sum_{i=1}^{n} \lvert \Delta w_{i1} \rvert$	0.9200	—	0.9115
$\sum_{i=1}^{n} \lvert \Delta w_{i2} \rvert$	—	0.9200	0.9115
τ_1	0.1178	—	0.2277
τ_2	—	0.1178	0.2277
τ^+	0.0433	0.0433	0.1770
τ^-	0.0746	0.0746	0.2776

	Account		
	#1	**#2**	**#1 + #2**
IR_1	0.28	—	0.2620
IR_2	—	0.27	0.2620
$\sum_{i=1}^{n} \lvert \Delta w_{i1} \rvert$	0.55	—	0.5492
$\sum_{i=1}^{n} \lvert \Delta w_{i2} \rvert$	—	0.92	0.9154
τ_1	0.04	—	0.1108
τ_2	—	0.12	0.15
τ^+	0.02	0.04	0.07
τ^-	0.03	0.07	0.15

	Account		
	#1	**#2**	**#1 + #2**
IR_1	0.2388	—	0.2174
IR_2	—	0.2759	0.2540
$\sum_{i=1}^{n} \lvert \Delta w_{i1} \rvert$	0.8948	—	0.8911
$\sum_{i=1}^{n} \lvert \Delta w_{i2} \rvert$	—	0.9196	0.8968
τ_1	0.1233	—	0.2225
τ_2	—	0.1148	0.2146
τ^+	0.0470	0.0430	0.1626
τ^-	0.0767	0.0750	0.2745

Constraints: $\bar{\sigma}_1 = \bar{\sigma}_2 = 5$ (top); $\bar{\sigma}_1 = 3$, $\bar{\sigma}_2 = 5$ (middle); $\bar{\sigma}_1 = \bar{\sigma}_2 = 5$, $w_{i,10} \leqslant 0$ (bottom).

client can be made better off without making another client worse off. But does Pareto optimality also ensure that all clients are treated fairly, ie, can we argue that the information ratios of two clients deteriorate at different speeds? In fact it is easy to see that there is no reason why information ratios should decline by the same percentage under nonlinear transaction costs. Any requirement like this would result into subsidising clients with limiting constraints (breadth of positions) at the expense of accounts that allow a broader universe. These accounts therefore have a much lesser need to engage into concentrated and transaction cost expensive portfolios.

We can apply the above analysis to investigate the capacity of a given investment management strategy. Arguably as we add an increasing number of accounts to a given strategy, transaction costs (market impact) will inevitably rise. Depending on the size and liquidity of a given market, the decay in information ratio can be substantial. To what degree that takes place can be investigated directly with the above methodology.[7]

8.8 REBALANCING PROBLEM

Suppose we are given a long-run strategic asset allocation that involves fixed weights on asset classes such as 25% US equities, 15% US bonds, 30% international equities, and so on. Differences between the performance of these asset categories will lead to deviations from targeted exposures, and investors will need to trade off the costs (transaction costs) and benefits (low tracking error) of any rebalancing program. Looking at it another way, you can have a hair cut every day and run up a huge bill at the barbers but enjoy the benefit of an optimal hair cut throughout the year. Alternatively, you can go to the barbers just once a year, in which case your bill will be low but the deviation from the optimal hair cut will be high.

We distinguish between trigger strategies, calendar-based strategies and tracking error-based strategies.

8.8.1 Trigger strategies

As before, we assume an investor that maximises utility of the form $u_{\mathrm{a}} - \frac{1}{2}\lambda\sigma_{\mathrm{a}}^2$, where μ_{a} and σ_{a}^2 denote active returns and active risk (squared tracking error). Suppose we look at rebalancing each asset individually against all other remaining assets in the portfolio. This is a very simplified version of the original rebalancing problem where all assets need to be rebalanced simultaneously depending on their individual deviations from the strategic asset allocation and the respective transaction costs. We also know that tracking error is linear homogeneous in weights (doubling weights will also double tracking error). This helps us to determine when to rebalance.[8] The benefit of rebalancing the ith asset class is given by $\frac{1}{2}\lambda\sigma_{\mathrm{a}i}^2\Delta w_i^2$, while the associated (linear) transaction costs are given by $\mathrm{tc}_i\Delta w_i^2$, where $\sigma_{\mathrm{a}i}^2$ describes active risk arising from the ith asset. We can define $\sigma_{\mathrm{a}i}^2$ more precisely as the risk associated with a long position in asset i, R_i, and a short position in the remaining portfolio, R_{-i}, ie

$$\sigma_{\mathrm{a}i}^2 = \sigma^2(R_i - R_{-i}) = \sigma_i^2 + \sigma_{-i}^2 - \sigma_i\sigma_{-i}\rho_{i,-i}$$

The optimal trigger point (deviation from the targeted weight) is the point where the net rebalancing benefit can be defined as

$$\tfrac{1}{2}\lambda\sigma_{\mathrm{a}i}^2\Delta w_i^2 - \mathrm{tc}_i\Delta w_i \tag{8.25}$$

Solving for Δw_i to arrive at a zero net balancing benefit, we get

$$\Delta w_i = \frac{\mathrm{tc}_i}{\frac{1}{2}\lambda\sigma_{\mathrm{a}i}^2} = \frac{\mathrm{tc}_i}{\frac{1}{2}\lambda(\sigma_i^2 + \sigma_{-i}^2 - \sigma_i\sigma_{-i}\rho_{i,-i})} \tag{8.26}$$

- High risk aversions will lead to smaller trigger points, reflecting investors who feel increasingly unhappy about deviations from the strategic asset allocation.
- High transaction costs will lead to larger trigger points as it becomes more costly to rebalance.
- Trigger points tend to be smaller for assets with high volatility and/or high covariation with the rest of the portfolio.

The next question we need to ask is where to rebalance to. Any reweighting will lead to a new net balancing benefit, as formulated in 8.25. Thus it is optimal to rebalance up to the point where the change in net rebalancing benefit is zero (ie, shortly before it becomes negative). We take the derivative of 8.25 with respect to Δw

$$\frac{\delta(\frac{1}{2}\lambda\sigma_{\mathrm{a}i}^2\Delta w_i^2 - \mathrm{tc}_i\Delta w_i)}{\delta\Delta w_i} = 0 \tag{8.27}$$

We reach this point at

$$\Delta w_i = \frac{\mathrm{tc}_i}{\lambda\sigma_{\mathrm{a}i}^2} \tag{8.28}$$

which is just half the solution in 8.26. Hence, optimal rebalancing will not rebalance all the way back to the strategic asset allocation but, rather, stop half way as transaction costs do not justify further rebalancing.

While at first glance intuitive, the above solution suffers from several severe problems:[9]

- Even after all trigger points have been defined, there is no control over total tracking error. Matters become worse when volatilities and correlations change, which will also require a change in trigger points. However, the attractiveness of trigger points comes from the ease with which they are implemented and supervised.
- The solution does not advise investors what to rebalance into. In fact it is relative transaction costs that matter, as selling an asset in the process of rebalancing always means buying another asset instead.
- The solution only allows a very simple form of transaction costs, namely, linear costs. To the extent that this is unrealistic, the resulting solution is misguided as the marginal costs of rebalancing will jump.

Trigger strategies will continue to be used in practical applications as the resulting trigger points are easy to police and enforce. Legally, there will be little argument about breaches of investment guidelines.

8.8.2 Calendar strategies

Calendar strategies rebalance portfolios at certain prespecified intervals, ie, daily, weekly, monthly, quarterly, etc. The rationale for calendar-based strategies in portfolio rebalancing is to capture mean-reversion in asset prices. Rebalancing to a given constant mix is, effectively, a concave investment strategy that outperforms in the absence of trends. The choice of calendar interval depends on how long one believes it takes mean-reversion to occur. The main problem with calendar strategies (apart from the fact that there is little empirical evidence for mean-reversion) is that there is no portfolio adjustment between calendar points. Any large market movements that significantly distort relative weights in mid-August will not trigger any rebalancing in a monthly rebalancing strategy. When rebalancing finally takes place at the beginning of September, large drifts – and, hence, excessive tracking error – might already have occurred.

8.8.3 Tracking error strategies

Both trigger and calendar strategies offer little control over *ex ante* tracking error (note that we can only control *ex ante* tracking error). However, it is tracking error, $\bar{\sigma}^2_{\mathrm{TE}}$, which measures the distance to the long-run desired asset allocation. Keeping this distance under control at minimal cost is the objective common to most investors. In fact this is identical to the index tracking problem. We need to solve an optimisation problem where, instead of maximising return, we aim to minimise transaction costs. Letting b_i denote the benchmark weight for the ith asset

$$\begin{aligned}
\min_{w_i, w_i^+, w_i^-} & \sum_{i=1}^{n} (\mathrm{tc}_i^+ w_i^+ + \mathrm{tc}_i^- w_i^-) \\
\bar{\sigma}^2_{\mathrm{TE}} &= \sum_i \sum_j (w_i - b_i)(w_j - b_j)\sigma_{ij} \\
1 &= \sum_{i=1}^{n} w_i + \sum_{i=1}^{n} (\mathrm{tc}_i^+ w_i^+ + \mathrm{tc}_i^- w_i^-) \\
w^i &= w_i^{\mathrm{Initial}} + w_i^+ - w_i^- \\
w_i &\geqslant 0 \\
w_i^+ &\geqslant 0 \\
w_i^- &\geqslant 0
\end{aligned} \tag{8.29}$$

In practice, we would check every day whether

$$\sum_i \sum_j (w_i - b_i)(w_j - b_j)\sigma_{ij} \geqslant \bar{\sigma}^2_{\mathrm{TE}}$$

in which case we would run Equation 8.29 to find the most cost-effective solution to establish the targeted tracking error. Hence, rebalancing takes place back to the target tracking error rather than to benchmark weights or trigger points. The advantage of this solution is that it explicitly takes relative transaction costs into account when deciding which assets to rebalance into and out of. It also keeps tracking error within the acceptable range, which allows direct control of risk. Ranges are exposure measures that do not allow for risk.

8.9 TRADE SCHEDULING

So far we have assumed that buy and sell transactions can be executed in one go (one period). However, this is rarely the case, or, as an anonymous Northfields client put it, “Your optimiser told me to sell five million shares but from whom?” This question motivates the field of trade scheduling, ie, how do we optimally break up a large trade into smaller trades and what are the involved trade-offs. Suppose we hold X shares that we want to liquidate by time T.[10] We divide time into N trading intervals, where $\tau = T/N$ denotes the length of a single trading interval. Each interval is indexed by k. The purpose of trade scheduling is to find an optimal trajectory $x_0, x_1, \dots, x_k, \dots, x_N$, where x_k is the number of shares (level) still held at the end of the kth trading interval. Obviously, we require the boundary conditions that $x_0 = X$ and $x_n = 0$. The inventory of stocks moves across time according to $x_k = x_{k-1} + n_k$, where n_k denotes the number of stocks traded (flow) in period k. Without trading we assume an arithmetic Brownian motion for the traded asset

$$S_k = S_{k-1} + \mu\tau + \tau^{1/2}\sigma\varepsilon_k \tag{8.30}$$

where μ denotes the assets expected (arithmetic) growth rate, ε_k stands for a draw from a standard normal random variable and σ reflects the instantaneous asset volatility. Note that while Equation 8.30 is a reasonable approximation to a geometric Brownian motion for very short time intervals (where $S_k - S_{k-1}$ is small), it suffers from the assumption of normally distributed disturbances since higher-frequency series (ie, those where asset changes are small and are therefore well approximated by an arithmetic Brownian motion) of financial asset returns exhibit considerable deviations from normality. In order to introduce the impact from trading we add both a temporary market impact and a permanent market impact. The permanent market impact arises from the negative or positive information effect associated with a particular trade (eg, Warren Buffet buys a particular stock, trading is takeover related, etc). The effect is assumed to be linear in the flow of our trading strategy, ie, we add $-\gamma n_k$ to Equation 8.30

$$S_k = S_{k-1} + \mu\tau + \tau^{1/2}\sigma\varepsilon_k - \gamma n_k \tag{8.31}$$

The temporary market impact arises from orders that empty the cumulative order book faster than new quotes arrive. Temporary impact is associated with liquidity provision within the trading period and is assumed to have vanished (ie, to have no further effect on the equilibrium price) when a new trading period arrives. If we denote by S_k^{trade} the price at which a trade actually occurs (after subtracting fixed costs of trading a as well as variable costs bn_k), we further change Equation 8.31 to

$$S_k^{\text{trade}} = S_{k-1} - a - bn_k\tau^{-1} \tag{8.32}$$

Finally, we define the costs of a trading strategy as the difference between the pre-trade value of our asset position and the costs of executing the trading schedule, ie

$$c = XS_0 - \sum_{k=1}^{N} n_k S_k^{\text{trade}} \tag{8.33}$$

Substituting Equations 8.30–8.32 into Equation 8.33 and applying the expectations operator (to yield expected costs, ie, negative wealth change) as well as the variance operator (to yield the variance of wealth changes as a measure of the inherent risk in a trading schedule), we compute

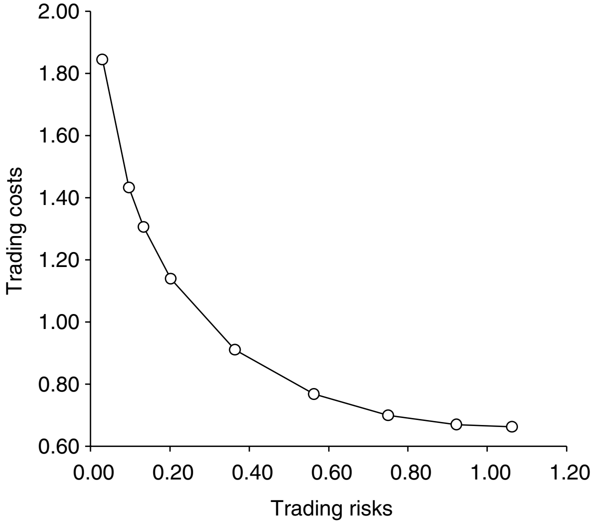
$$E(c \mid x_0, x_1, \dots, x_n) = -\mu\tau \sum_{k=1}^{N} x_k + \tfrac{1}{2}\gamma X^2 + aX + (b - \tfrac{1}{2}\tau\gamma) \sum_{k=1}^{N} n_k^2 \tag{8.34}$$

$$\operatorname{var}(c \mid x_0, x_1, \dots, x_n) = \tau\sigma^2 \sum_{k=1}^{N} x_k^2 \tag{8.35}$$

Why are we allowed to do this? Because we implicitly assume that our holding trajectory $x = (x_0, x_1, \dots, x_n)$ is fixed before we start trading. In other words it is deterministic, ie, there is no uncertainty around x_k when we apply the variance operator on Equation 8.33. Our optimal policy is not reactive, ie, there is no mechanism for "deviation trading", where we would speed up or slow down the execution as we move away from the benchmark price. This not only facilitates the derivation of Equations 8.34 and 8.35 but also allows us to easily solve for the optimal trade schedule, as we will see in the next section.

Assuming mean variance preferences, the optimal trading schedule can be derived by finding the optimal trade-off between expected costs of trading and its riskiness. This describes the trader's dilemma. A trader could execute immediately at very high costs arising from large temporary market impact, but avoid all uncertainty that stems from future but not yet known transaction prices.

Figure 8.3 Efficient trading frontier



Solution to Equation 8.36 and the associated costs and risks given in Equations 8.34 and 8.35 for our assumed parameterisation.

Alternatively, they could break up the trade into equal sizes to minimise the effect of temporary market impact, but this would come at the expense of significant execution risk (from exposing the trading strategy to the possibility of large adverse price moves). For a given risk-aversion parameter λ we simply solve

$$\min_{x} E(c \mid x) + \lambda \operatorname{var}(c \mid x) \tag{8.36}$$

Note that while some have worked on elegant closed-form solutions, this is not necessary and can prove restrictive in some cases. All we need to do is to numerically find the solution to Equation 8.36 with our optimiser of choice by choosing $x_0, x_1, \dots, x_n$ under the constraint that $x_0 = X, x_N = 0, x_k = x_{k-1} + n_k$. This can easily be done even on an Excel spreadsheet. We could also add additional constraints as limits on n_k, stepwise linear functions for temporary market impact, etc.

Varying λ, we arrive at the efficient trading frontier (ETF), ie, the set of all trading strategies that offer the lowest trading costs for a given level of trading risks. We assume $S_0 = 50$, $X = 100{,}000$, $T = 5$, $N = 5$, $\tau = 1$, $\sigma = 0.95$, $\mu = 0$, $a = 0.0625$, $\gamma = 2.5 \times 10^{-7}$, $b = 0.0000025$ and plot the ETF in Figure 8.3. Once the ETF

Table 8.7 Optimal trading trajectories associated with the efficient trading frontier

	$\lambda/10^6$								
	10	4	3	2	1	0.5	0.25	0.1	0
$x_0 = 10^6$	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
$x_1/10^5$	1.78	3.12	3.60	4.29	5.42	6.37	7.06	7.58	8
$x_2/10^5$	0.32	0.97	1.29	1.83	2.90	3.95	4.78	5.45	6
$x_3/10^5$	0.06	0.30	0.46	0.77	1.48	2.28	2.95	3.52	4
$x_4/10^5$	0.01	0.09	0.15	0.28	0.62	1.04	1.41	1.73	2
$x_5 = 0$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$E(c \mid x)/10^6$	0.03	0.10	0.13	0.20	0.36	0.56	0.75	0.92	1.06
$\operatorname{var}(c \mid x)/10^{12}$	1.84	1.43	1.31	1.14	0.91	0.77	0.70	0.67	0.66

This table provides solutions $x_0, x_1, \dots, x_n$ along the efficient trading frontier in Figure 8.3.

is derived we can apply all the tools we know from mean variance investing. Let us introduce the principal bid as a risk-free trading strategy. If we assume the principal bid to come at a cost of 1.4 and, by definition, zero risk, this would add a point on the y-axis in Figure 8.3 analogous to the risk-free rate in portfolio theory. Having established a risk-free strategy, it is then trivial to conclude that a combination of the risk-free trading strategy and a mean variance efficient trading strategy will reduce trading costs at the same level of risk. We can also define the tangency portfolio (the point on the ETF where a ray through the principal bid becomes a tangent to the ETF) as the strategy that has the highest likelihood of offering lower costs than the principal bid, etc. How do the different trading schedules underlying the ETF vary with risk aversion? We can infer this from the associated trading schedules in Table 8.7. Risk-neutral traders ($\lambda = 0$) will sell (or buy) the required amount of stocks in equal instalments in order to minimise temporary market impact, without any concern for the associated risks. As traders become more risk sensitive, the trading schedule accelerates, ie, more stocks are sold during the first periods, pushing up the total cost but reducing risks.

Almgren and Chriss (2000) also extend the above single-stock framework to a portfolio of stocks. This allows them to find the optimal trading strategy in a more realistic setting. Suppose a manager needs to rebalance their portfolio. We can view the trade list as a long–short portfolio, where long stocks have negative drift (alpha) as we want to sell them and short stocks have positive drift (alpha) as we want to buy them. The covariance matrix now also plays a role in deriving the trading schedule. Long–short positions of highly correlated stocks expose the trader to little risk, while it becomes very risky to leave uncorrelated stocks untraded.

8.10 OPTION REPLICATING STRATEGIES

8.10.1 Portfolio insurance

Portfolio insurance (PI) strategies are, essentially, option replicating strategies. In theory they require continuous trading, and as such create infinite transaction costs. In its simplest form PI provides investors with upside participation (always less than 100%) in risky assets (for example, equities), at the same time guaranteeing a minimum return (always less than the risk-free rate) over a prespecified time horizon. In other words, the investor gives up some downside in exchange for a guaranteed return.

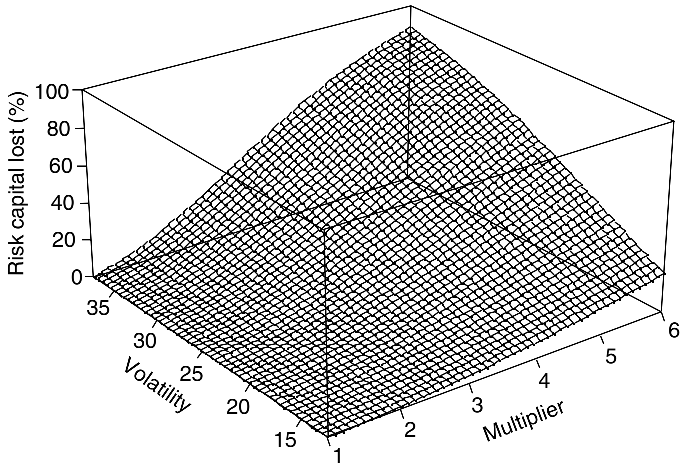
Obviously PI involves optionalities as the investor enjoys the best of equities and cash (minus the cost of this option). High volatility must increase the value of an investor's right to achieve the best return *ex post*, while low interest rates will increase the amount of assets that need to be held in riskless securities in order to satisfy the guaranteed return. High volatility increases optionalities and therefore reduces participation in the risky asset. However, relative to a low-volatility environment, high volatility also raises the odds of experiencing large positive and negative returns, which must exactly offset the volatility costs. It is true that low rates force insurance providers to set aside large amounts of capital to ensure that guarantees are met, but it is equally true that low rates reduce the opportunity costs of not investing in competing investments. In the misguided hunt to offer appealing insurance products, suppliers of PI reduce protection levels, change payouts (for example, participation in an average index rather than an end-of-period index), take on tail risks (writing out-of-the-money puts, digitals, etc) or change the underlying (participation in a price index instead of a performance index). PI can be generated either by buying or by replicating the associated options. Direct replication by investors is often preferable as it does not expose the investor to intellectual risk (option profits and losses are viewed on a standalone basis rather than in a portfolio context and so are associated with reckless speculation), nor do they generate excess profits for investment banks. PI deliberately generates positively skewed distributions by transforming a low probability of large losses into a high probability of small returns. By their very construction insured portfolios offer attractive average returns (generated by a few very large returns) with poor median returns (the level below which 50% of all returns fall). However, the median is a much better measure of location if returns are not distributed symmetrically.

8.10.2 Rebalancing rules are key

PI is often implemented by replicating option-like payouts using dynamic hedging. The most common approach to this is the so-called CPPI strategy (constant proportion PI). It is driven by a simple allocation rule that provides investment (entry strategy) and disinvestment (exit strategy) decisions. The amount of investment into equities is calculated as the product of a multiplier and a "cushion". The cushion is defined as the difference between the current value of the CPPI portfolio (for example, 100) and the present value of the targeted floor (for example, for a 4% discount rate the present value of 90 in three years' time is 80). We can interpret the cushion as risk capital. The multiplier translates changes in risk capital into changes in allocation. For example, a multiplier of 3 will lead to a 30% equity allocation in the above example, while the remaining 70% is invested in the riskless asset (zero bonds with three years' maturity). If the cushion increases from 10 to 12, we arrive at a 36% equity allocation. Let us ignore interest rates for a moment and assume that equities fall (instantaneously) by 33%. Our risky assets fall from 30 to 20, but we are still left with 70 in the riskless asset. The total portfolio value declines to 90, which equals the targeted floor. It is not a coincidence that one over the multiplier equals 33%.

CPPI provides investors with a stochastic guarantee. If markets drop instantaneously – or over any time period where we cannot readjust the portfolio allocation (typically overnight, but it could also happen over the course of a week if exchanges are closed, as on 9/11) by more than the inverse of the multiplier – a CPPI strategy will fail to

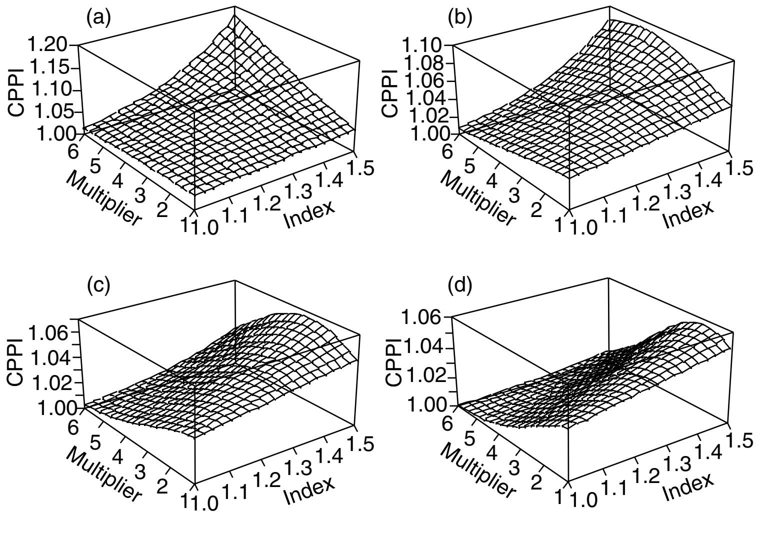
Figure 8.4 Volatility cost



protect the floor. Note that a gradual loss of 33%, or even much more accompanied by readjustments, does not pose a problem. As CPPI is effectively a trend-following strategy, it will buy (sell) equities after they have risen (fallen). Buying high and selling low will lose money. However, these costs are close to the costs of an equivalent option (in which case the option provider runs a similar strategy). Note that it is *ex post* realised volatility that determines the volatility costs rather than *ex ante* expected volatility. If volatility is higher than expected, a CPPI strategy will result in lower than expected returns. Large multipliers will result in high turnover, as well as high volatility costs. A useful measure of the inherent costs of a CPPI strategy is to ask what fraction of the provided risk capital (cushion) would be lost if equities returned the same as riskless bonds. This relationship is plotted in Figure 8.4 (one-year CPPI with 25% volatility on EuroStoxx 50).

Large multipliers result in substantial volatility costs (up to 90% of the cushion). Also note that volatility costs for large multipliers triple if the volatility rises from 15% to 35%. Multipliers are principally derived from historical data (worst-case historical returns) or from volatility estimates (worst-case return at a prespecified confidence level from a volatility forecasting model or Extreme Value Theory). If forecasts signal rising volatility, equity allocations might be reduced even if the portfolio value has risen. It can be shown that CPPI strategies are a special form of value-at-risk-based strategies.

Figure 8.5 The effect of volatility and multiplier on CPPI performance



(a) 20% volatility; (b) 30% volatility; (c) 40% volatility; (d) 50% volatility.

The obvious risk is gap risk, ie, the risk that equities fall by more than the inverse of the multiplier. However, there are two more subtle, but equally important, risks. The first is interest rate risk. Usually if equity markets decline in a crisis scenario, interest rates will be pushed down by safe haven arguments. However, this lowers the rate at which the disinvested equity (equity sold during a market crash) can be reinvested. The second often overlooked risk is the sensitivity to realised volatility.

Let us evaluate a CPPI with a one-year time horizon and a floor of 100. Interest rates are 3.5%. Figure 8.5 shows the expected participation (z-axis) of a CPPI strategy as a function of the multiplier (y-axis) and the underlying equity index performance (x-axis). Although large multipliers deliver the best performance at modest volatility levels, their performance is disastrous for high volatility levels, where they deliver the floor even when equity performance is high.

8.10.3 Trading rules and turnover reduction

By definition it is difficult to improve on a forecast-free, transparent and rule-based strategy (and hence passive) without altering the concept. We aim to limit turnover without compromising performance or taking on larger risks. Traditionally, rebalancing rules aim to adjust portfolio weights back to target weights (set by a CPPI formula) as soon as they fall outside a prespecified corridor. However, marginal rebalancing back to the upper and lower barriers of the no-trading corridor considerably reduces trading costs, as demonstrated in Table 8.8.

For a 10% corridor (meaning that if the target allocation is 5% we trade if weights are above 5.5% or below 4.5%) our rebalancing rule saves about 50% in turnover for a multiplier of 4. Marginal rebalancing reduces turnover at the boundaries (unless triggering is very frequent) relative to central rebalancing, while at the same time maintaining the opportunity to move back inside in response to favourable relative movements. We might suspect that risks will also rise as marginal rebalancing forces us to stay longer at the boundaries and thus face increased gap risk.

Table 8.9, which is based on a large number of Monte Carlo simulations, shows the contrary. For modest multipliers there is no increased gap risk at all. Investors will only experience slightly higher risks if multipliers become large (above 8). This methodology proves to be extremely valuable when underlyings are illiquid or if futures are not available. Note that the exact corridor is a function of volatility and transaction costs and needs to be calibrated for different risky assets.

8.10.4 Who should buy PI?

Countless books on option pricing offer numerical guidance on the evaluation of even the most complicated exotic options but remain surprisingly silent on the question, who should buy PI?[11] Although everyone benefits from the pooling of independent insurance risks, engaging in PI is a zero sum game. For every buyer of PI there must be a seller. So we need to establish criteria to help us determine what type of investor should buy (or sell) PI. Let us assume that investors maximise expected utility from end-of-period wealth and exhibit decreasing marginal utility from wealth. Under these innocent assumptions, the finance literature makes three statements about PI.

- Investors whose risk aversion falls faster (with increasing wealth) than the average risk aversion (risk aversion of the market) will demand convex strategies (PI). Those strategies translate rising wealth levels into larger equity allocations.

Table 8.8 Turnover reduction (%)

(a) Naive strategy

	Trigger points (%)					
Multiplier	**0**	**5**	**10**	**15**	**20**	**25**
2	35	16	9	6	4	3
4	189	144	106	81	64	51
6	322	255	220	179	151	125
8	380	287	263	237	204	179
10	375	300	270	250	219	195

(b) DeAM strategy

	Trigger points (%)					
Multiplier	**0**	**5**	**10**	**15**	**20**	**25**
2	35	8	4	3	2	1
4	189	93	59	42	31	23
6	322	201	140	105	81	64
8	380	266	196	150	119	96
10	375	283	217	171	138	114

Equally, falling wealth levels will be accompanied by reduced equity allocations. Note that we need to compare individual risk aversion to average market risk aversion as we need to evaluate the decision to buy PI within market equilibrium.

- Investors with above-average expectations (higher return expectations than the market) will also invest into PI. Intuitively this makes sense as it is well known that PI will only provide investors with significant returns if realised returns on the underlying stock market are considerably higher than average returns. In other words, investors with average risk aversion need to be optimistic about the market. This is not surprising as optimistic investors tend to invest more aggressively and therefore need to protect their downside.
- All path-dependent strategies should be avoided by the above types of investor as every path-dependent strategy is dominated by a corresponding (with the same expected return)

Table 8.9 Rebalancing and gap risk (%)

(a) Naive strategy

	Trigger points (%)					
Multiplier	**0**	**5**	**10**	**15**	**20**	**25**
2	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.20	0.00	0.20
10	1.20	1.20	1.60	2.60	3.20	3.80

(b) DeAM strategy

	Trigger points (%)					
Multiplier	**0**	**5**	**10**	**15**	**20**	**25**
2	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0
8	0.0	0.0	0.2	0.2	0.2	0.2
10	1.4	1.6	2.4	3.5	4.4	5.3

path-independent strategy. Path dependency creates additional volatility of end-of-period wealth without increasing expected return.

Equipped with the above analysis, we can evaluate popular PI strategies. Traditional PI such as protected put and CPPI are ideally suited for investors with a relative risk aversion that decreases as their rebalancing rules translate rising wealth levels into increased equity weightings. Pension funds are a natural group of investors whose risk aversion increases faster with falling wealth than the average risk aversion in the market. The smaller a pension fund surplus becomes, the larger the risk aversion. Insurance products that lock in intermediate wealth levels (ratcheting) or guarantee past highs destroy value for long-term investors as intermediate wealth obviously has no value for investors concerned about terminal wealth. In fact investors throw away considerable amounts of money if they buy these products! Discount certificates (effectively covered call writing) translate rising wealth levels into falling equity allocations, while falling wealth levels lead to rising equity allocations. To find them appealing, investors need to have an increased willingness to take on extra risk if their wealth level is falling. It is not clear why retail investors and private clients (target clients for this kind of product) who are described as conservative would want to invest in strategies that generate large losses when they need their money most (as their marginal utility from consumption will be high if their wealth level is low). Note that this also applies to writing deep out-of-the-money puts to partially finance PI.

At this point it is worth bringing to the attention of investors some additional features of PI that are often overlooked in practice.

- PI typically applies to very liquid asset classes (preferably with liquid future markets) as it makes hedging easier for the insurance provider. For long-term investors this means that they are giving up diversification and lose the liquidity premium. The same is true for insurance programs on illiquid asset classes (like hedge funds). In this case the liquidity premium (and more in what are often quite untransparent transactions) is paid to the liquidity provider (insurance seller).
- Investors often apply PI to a subset of their equity holdings. This generates overhedging. An insurance program on foreign equities will hedge out risks that it is not necessary to hedge as the correlation between foreign and domestic equities is less than one.
- The single-period returns of insured portfolios exhibit the deliberately engineered significant (positive) skewness that makes the concept attractive to the investors identified above. However, PI is often periodically reset (rolled over). Unfortunately, the Central Limit Theorem tells us the product of independent and identically distributed returns will become approximately lognormal after about 30 draws. Even though single-period return distributions exhibit significant skew, repeated investing into reset strategies will lose this property in the long run.

Repeatedly investing into annual PI on liquid asset classes on a subset of assets will lead to an underdiversified and overhedged portfolio that also loses its skewness in the long term.

Pension funds have been made painfully aware of the risks associated with their often large equity exposure. As a consequence, traditional PI (protective put, CPPI) has proved to be very popular. But is this the correct answer? Recall that PI provides the investor with the maximum of equities or cash (minus premium). Falling equity markets will result in a synthetic cash position. Cash, however, is one of the most risky assets in an asset liability context as it carries zero duration. Assuming a 30-year duration on liabilities, a pension fund that is invested into 50% equities will see a 15% erosion of its surplus from equities alone if rates drop by 100bp. What pension funds really need is an option (insurance program) that provides the best of equities and long bonds.

Liability-driven investors will find that traditional PI still exposes them to large risks as holding cash proves to be risky for investors with long-term liabilities.

8.10.5 Forecast-free, or not quite?

Equity-protection strategies are often advertised as being forecast-free. We have already seen above that this is not the case. Investors who exhibit average (market) risk aversion will only buy PI when they are more bullish than the market consensus. Quantitative analysts are (very often to their frustration) asked to come up with an "optimal" protection product. However, without knowledge of an investor's utility there is no optimal product. The practical answer is that the art in selling insurance programs (or options) is to set up structures which, in a way, take risks that are not perceived to be risky. One obvious example is to repeatedly sell deep out-of-the-money puts to finance puts bought at higher strikes. While most of the time this strategy will make PI cheaper, there is one scenario (when equity returns are sufficiently negative) in which all gains are returned. All this strategy does is to bet against an infrequent event that might not have been realised in the past and therefore looks riskless to many investors.

What has been said above can also be turned around. Every concept has some scenarios in which it performs better than other concepts. Investors need to make sure that these scenarios are the ones that are relevant to them. Suppose you consider investing into either a CPPI strategy or a protective put. Hedging costs are paid up front in a protective put strategy and depend on *ex ante* expected volatility. For a CPPI strategy, however, costs depend on *ex post* realised volatility. Investors expecting realised volatility to be lower than *ex ante* expected volatility will prefer CPPI-type (replicating) strategies. Alternatively, some investors might want to minimise the regret of investing in the wrong asset class. These investors want to buy an option (or hedging program) to achieve the maximum of equities and bonds (less premium). If the performance difference between the two asset classes is large enough, it can cover the costs of providing the option and still outperform a simple balanced portfolio. If the relative performance difference is small, investors will experience underperformance relative to a balanced fund. Compared to CPPI, this type of strategy does not protect your wealth as investors lose out if both assets classes fall by 20% (the loss amounts to 20% plus the additional loss of premium).

EXERCISES

1. Replicate Figure 8.3 and Table 8.7 using numerical optimisation in Excel.

2. Assume the principal bid strategy comes at a cost of 1.4×10^6. Find the tangent to the ETF in Figure 8.3 and calculate $\text{prob}(c \leqslant 1.4 \times 10^6)$ along the ETF. Comment on your findings.

[1] See Sharpe (1991).

[2] Grinold and Kahn (2000, p. 452, Equation 17.4).

[3] In matrix form we write $\min \boldsymbol{w}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{w}, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu} = \mu, \boldsymbol{w}^{\mathrm{T}}\mathbf{1} = 1, \boldsymbol{w} = \boldsymbol{w}^{\text{Initial}} + \boldsymbol{w}^{+} - \boldsymbol{w}^{-}, \boldsymbol{w}^{+} - \boldsymbol{w}^{-} \leqslant \boldsymbol{\tau}$, $\boldsymbol{w}, \boldsymbol{w}^{+}, \boldsymbol{w}^{-} \geqslant 0$.

[4] Inclusion of transaction costs can be found in Mitchell and Braun (2002) or Lobo *et al* (2002).

[5] Transaction costs for asset purchases would then need to be distinguished by

$$\tau_i^{+} = \theta_i^{+}(\Delta w_i^{+})^{\gamma_i^{+}} \quad \text{for all } i = 1, \dots, n$$

This, however, would complicate 8.16 without additional insights.

[6] We have used NuOPT for S-Plus (Version 1.0) for the following numerical examples.

[7] A more detailed treatment of this problem can be found in O'Cinneide *et al* (2006).

[8] This section follows Masters (2003).

[9] See Donohue and Yip (2003) for the correct solution.

[10] This section will make extensive use of the trade scheduling framework laid out by Almgren and Chriss (2000). Interested readers should refer to this for more details.

[11] This section follows Leland (1988).

REFERENCES

Almgren, R., and N. Chriss, 2000, "Optimal Execution of Portfolio Transactions", *Journal of Risk* 3, pp. 5–39.

Donohue, C., and K. Yip, 2003, "Optimal Portfolio Rebalancing with Transaction Costs", *Journal of Portfolio Management* 29, pp. 49–63.

Grinold, R., and R. Kahn, 2000, *Active Portfolio Management*, Second Edition (New York: McGraw-Hill).

Leland, H., 1988, "Who Should Buy Portfolio Insurance?", in D. L. Luskin (ed), *Portfolio Insurance: A Guide to Dynamic Hedging* (John Wiley & Sons).

Lobo, M., M. Fazel and S. Boyd, 2002, "Portfolio Optimisation with Linear and Fixed Transaction Costs", Mimeo.

Mitchell, J., and S. Braun, 2002, "Rebalancing an Investment Portfolio in the Presence of Transaction Costs", Mimeo.

Masters, S. J., 2003, "Rebalancing", *Journal of Portfolio Management* 29, pp. 52–7.

O'Cinneide, C., B. Scherer and X. Xiaodong, 2006, "Ensuring Fairness when Pooling Trades", *Journal of Portfolio Management*, pp. 33–43.

Sharpe, W. F., 1991, "The Arithmetic of Active Management", *Financial Analysts Journal* 47, pp. 7–9.

9

Portfolio Optimisation with Options: From the Static Replication of CPPI Strategies to a More General Framework

9.1 INTRODUCTION

We finished the last chapter with constant proportion portfolio insurance. This is a well-known example of a continuous trading strategy and its properties have been studied extensively in the literature.[1] While the continuous time framework is predominant in modern finance it is at best an approximation to reality. Even if continuous trading was feasible, the presence of transaction costs would make continuous trading infinitely costly as the sum of absolute stock price increments becomes infinite. Transaction costs will therefore enforce discretisation, ie, the investor rebalances only a finite number of times in an attempt to trade off transaction costs against replication error. Various trading policies to improve this trade-off have been introduced. Instead we want to find a static buy and hold strategy of a few traded options that approximates the continuous (and realistically unattainable) CPPI trading strategy as closely as possible. The static option hedge would incur no trading cost and requires no rebalancing. It also takes away the arbitrariness of simulation based comparisons. While the tracking criterion (sum of probability weighted squared deviations) is defined under the real-world distribution, ie, the P measure, the available options are priced under the risk-neutral distribution, ie, the Q measure.

We will show that our approach to portfolio optimisation can in general be used to include options into portfolio allocation decisions. The trick is to make the risk-neutral Q measure consistent with the real-world P measure, ie, we derive the RND from observed option prices (Q measure), derive from it a forward-looking real-world distribution (P measure) and glue the marginal distributions together using copula functions (start from Gaussian). This has four main advantages.

- It is a forward-looking distribution and you can use the information about the future distribution (volatility, degree of non-normality) contained in today's option prices.
- Option prices are consistent with asset prices, ie, the scenarios you use for asset returns are also the scenarios that have been used for option pricing.
- You can optimise utility under the P measure subject to a budget constraint under the Q measure to look at different option strategies and how they affect utility as well as how you can game traditional performance measures (Sharpe, Sortino, Omega, etc). Note that optimisation using options is not so easy as you cannot use the strike as a variable: a nonlinear equality (budget) constraint leads to a non-convex problem. Use mixed integer programming instead (many strikes, but use only n of them) is also a non-convex but there are established routines.
- We can look whether non-normality in the marginal distributions is more or less important than tail dependence (which you control through your copula choice).

In general it is the authors' belief that we will see that many techniques originally applied in the option pricing arena will shift to be applied in portfolio construction problems.

9.2 THE TRACKING PROBLEM

In this section we will show that tracking a CPPI strategy is a straightforward application of the martingale approach to portfolio optimisation that can in general be used to include options into portfolio allocation decisions.[2]

Suppose an investor wants to track a particular payout profile $W^*_{T,i}$ that has been generated by an arbitrary dynamic trading strategy across $i = 1, \dots, n$ states of the world at time T. In the special case of a CPPI (constant proportion portfolio insurance) we know that the end of period wealth strategy can be described in closed form as[3]

$$W^*_{T,i} = F_T + (W_0 - F_0)\left(\frac{S_{T,i}}{S_0}\right)^{\alpha} \exp\{(r - \alpha(r - \tfrac{1}{2}\sigma^2) - \tfrac{1}{2}\alpha^2\sigma^2)T\} \quad (9.1)$$

where F_T denotes the accumulated floor that started at F_0 and has since grown at the risk-free rate r, W_0 stands for the starting period wealth, $S_{T,i}/S_0$ describes on plus the total return on an investment in the underlying risky asset for scenario i, α is called the multiplier that determines how fast a dynamic process rebalances in and out the risky asset and σ describes the annual volatility of the underlying risky asset (log) returns. However, for Equation 9.1 to hold exactly we need the assumption of continuous trading in frictionless markets without transaction costs. In reality this path independent Nirwana solution is never achieved and the existence of transaction costs will make the final payout path dependent. In the following we will therefore outline an efficient procedure to track a given CPPI with an options portfolio. In order to evaluate the closeness (tracking quality) of a given static investment strategy we suggest a mean squared error penalty function

$$\min_{w_{\mathrm{C}}, w_{\mathrm{S}}, w_{\mathrm{C}_1}, \ldots, w_{\mathrm{C}_m}} \sum_{i=1}^{n} \pi_i (W^*_{T,i} - W_{T,i})^2 \tag{9.2}$$

where $w_{\mathrm{C}}, w_{\mathrm{S}}, w_{\mathrm{C}_1}, \ldots, w_{\mathrm{C}_m}$ stand for the number of units invested into cash, stocks, as well as m call options with respective strikes X_m. For a given scenario i each option pays off an amount

$$C_{m,i} = (S_{T,i} - X_m)^+ = \max(S_{T,i} - X_m, 0)$$

at time T. Hence, the end of period wealth (at time T) for a static buy and hold portfolio in state i is given by

$$W_{T,i} = w_{\mathrm{C}} \mathrm{e}^{rT} + w_{\mathrm{S}} S_{T,i} + w_{\mathrm{C}_1} C_{1,i} + \cdots + w_{\mathrm{C}_m} C_{m,i} \tag{9.3}$$

Note that we only include calls as we can easily replicate puts out of equities, calls and cash using the put-call parity. However, given that we can only invest W_0 we need to ensure that all call options are priced correctly with respect to the assumed scenarios. In other words the expected value of future payouts under the risk-neutral distribution needs to amount to current wealth

$$W_0 = \mathrm{e}^{-rT} E_Q(W_T) = \mathrm{e}^{-rT} \sum_{i=1}^{n} \bar{\pi}_i W_{T,i} \tag{9.4}$$

where $E_q(W_T)$ denotes the expected value of W_T across all i scenarios, where each scenario is weighted according to the risk-neutral probabilities, $\bar{\pi}_i$. Equation 9.4 essentially operates as a budget constraint that ensures that all options used in the optimal tracking problem are properly priced and that no arbitrage opportunities exist as by definition

$$\mathrm{e}^{-rT}\sum_{i=1}^{n}\bar{\pi}_i W_{T,i}^{*}$$

also must equal W_0. Note the distance in Equation 9.2 can be made arbitrarily small as we increase the number of options involved.

The budget constraint (Equation 9.4) is highly nonlinear in strike prices X_m. We know from optimisation theory, that the feasible set of an optimisation problem involving nonlinear equality constraints is not a convex set, irrespective whether the equality constraint itself is convex or not. To avoid the computational challenges associated with non-convex optimisation we assume a grid of strike prices and solve Equations 9.2–9.4 for a limited number of options using well developed mixed integer programming technology.[4] In other words, instead of having to choose X_m directly we offer the optimiser a whole set of strikes of which it is only allowed to take a limited number. This is often described as imposing a cardinality constraint. Fixing the strikes transfers the nonlinear budget constraint in a simple linear equality constraint. While this comes at the expense of using mixed integer variables (fixed number of options) which themselves create non-convexity (as it involves discrete steps), well established branch and bound algorithms exist for both linear and nonlinear objective functions.

Alternatively we could do a "manual" search. For a given number of assets n, we suggest solving a series of sub-problems with a limited universe of n_{assets}. However, in total there will be

$$\frac{n!}{(n-n_{\text{assets}})!}\frac{1}{n_{\text{assets}}!}$$

sub-problems. For 13 assets there will be 1,716 different combinations of six assets. We therefore suggest the use of mixed integer variables using a cardinality constraint instead.

9.3 DERIVING THE *Q* AND *P* MEASURE

In order to implement the above approach we need to find both risk-neutral ($\bar{\pi}_i$) as well as real-world probabilities (π_i) that are mutually consistent across all scenarios (i). This has long been perceived as a major hurdle. "While the state-preference approach is perhaps more general than the mean variance approach and provides an elegant framework for investigating theoretical issues, it is unfortunately difficult to give it empirical content".[5] We start in a world, where returns are lognormal distributed and volatility is constant. In this textbook case we can calculate the price of state price security (p_i) using

$$p_i = \mathrm{e}^{-rT}\biggl(N\biggl(\frac{\ln(S_0/S_i) + (r - d - 0.5\sigma^2)^{\mathrm{T}}}{\sigma\sqrt{T}}\biggr) - N\biggl(\frac{\ln(S_0/S_{i+1}) + (r - d - 0.5\sigma^2)^{\mathrm{T}}}{\sigma\sqrt{T}}\biggr)\biggr) \tag{9.5}$$

Equation 9.5 shows the price of a security that pays off one monetary unit if the stock price falls between S_i and S_{i+1} and zero otherwise. We use standard notation where S_0 is the current stock price, r, d are risk-free rate and dividend yield, σ denotes volatility and T means time to maturity. Essentially this is a long–short portfolio of two standard digital options.[6] Let us illustrate this with an example. Suppose $S_0 = 6{,}000$, $r = 0.04$, $d = 0$, $T = 1$, $\sigma = 0.2$. The calculations are summarised for a step size of 250. As a portfolio of all state price securities replicates the risk-free asset, we know that

$$\sum_i p_i \cdot 1 = \mathrm{e}^{-rT}$$

This can be restated in terms of so-called risk-neutral probabilities

$$\sum_i \bar{\pi}_i = 1$$

where $\bar{\pi}_i = p_i \cdot \mathrm{e}^{-rT}$. The last two columns of Table 9.1 provide the scenario "midpoint" as well as the risk-neutral expectation.

How can we check our results? For a start we know that the sum of all risk-neutral probabilities must amount to 1 and that the sum of the state prices equal the value of the risk-free asset. Using the numbers in Table 9.1 we see this is indeed the case

$$\begin{aligned}\sum_i \bar{\pi}_i &= \cdots + 5.35\% + 6.55\% + \cdots + 5.25\% + 4.38\% + \cdots \\ &= 1\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} p_i &= 0.96 \approx \mathrm{e}^{-rT} \\ &= 0.96\end{aligned}$$

Table 9.1 Derivation of state prices (delta securities) in a Black–Scholes world

S_i	S_{i+1}	P_i (%)	$\bar{\pi}_i$ (%)	$(S_i+S_{i+1})/2$	$\bar{\pi}_i(S_i+S_{i+2})/2$
$\vdots$					
4,750	5,000	5.14	5.35	4,875	260.7
5,000	5,250	6.29	6.55	5,125	335.6
5,250	5,500	7.20	7.50	5,375	402.9
5,500	5,750	7.77	8.09	5,625	455.1
5,750	6,000	7.97	8.30	5,875	487.4
6,000	6,250	7.81	8.13	6,125	497.9
6,250	6,500	7.35	7.65	6,375	487.9
6,500	6,750	6.68	6.96	6,625	460.8
6,750	7,000	5.89	6.13	6,875	421.2
7,000	7,250	5.04	5.25	7,125	373.9
7,250	7,500	4.21	4.38	7,375	323.3
$\vdots$					

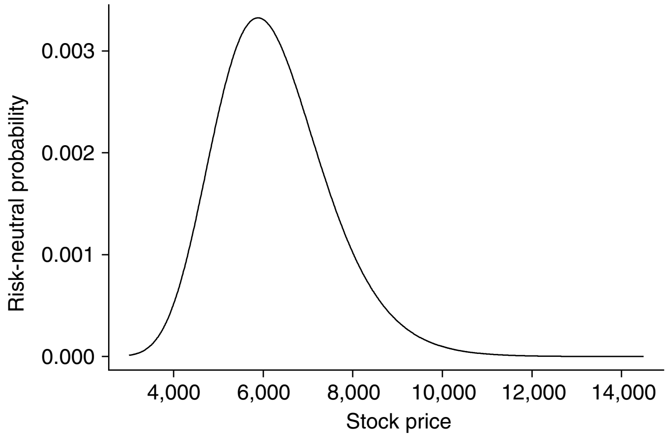
State prices are calculated according to Equation 9.5 using $S_0 = 6{,}000$, $r = 0.04$, $d = 0$, $T = 1$, $s = 0.2$.

Next we check, whether our risk-neutral probabilities correctly price our risky asset that trades at $S_0 = 6{,}000$. To do so we calculate the discounted expected value of one year stock prices under the risk-neutral distribution

$$\mathrm{e}^{-rT} E_Q\left(\frac{S_i + S_{i+1}}{2}\right) = \mathrm{e}^{-rT} \sum_i \bar{\pi}_i \frac{S_i + S_{i+1}}{2} = 6{,}000 = S_0$$

As we expected our pricing model recovers the current stock price. Decreasing the step size to 10 we arrive at the risk-neutral distribution in Figure 9.1, which resembles a continuous lognormal distribution. Finally we need to calculate (better calibrate) the set of real-world probabilities that is consistent with our risk-neutral distribution. Let us express the real-world probability as a function of risk aversion as well as risk-neutral probabilities. We know from elementary asset pricing theory that the state price deflator, Λ_i, for state i equals the ratio of risk-neutral to real-world probability, divided by one plus the risk-free rate: $\Lambda_i = \mathrm{e}^{-rT}\bar{\pi}_i/\pi_i$. At the same time we know that the state price deflator equals the marginal utility of consumption times an arbitrary constant $\Lambda_i = \theta W_{T,i}^{-\gamma}$.[7] Equating both

Figure 9.1 Risk-neutral distribution implied by state price approximation



Risk-neutral probabilities are calculated using Equation 9.5 with $S_0 = 6{,}000$, $r = 0.04$, $d = 0$, $T = 1$, $\sigma = 0.2$ for a step size of 10 and multiplying the state price by e^{rT}.

expressions and solving for π_i yields

$$\pi_i = \frac{e^{-rT} \cdot \bar{\pi}_i \cdot W_{T,i}^{\gamma}}{\theta} \tag{9.6}$$

In order for $\sum_i \pi_i = 1$ we need to determine θ which is given by

$$\theta = e^{-rT} \sum_i \bar{\pi}_i \cdot W_{T,i}^{\gamma}$$

Substituting this back into Equation 9.6 provides us with the required relationship

$$\pi_i = \frac{e^{-rT} \cdot \bar{\pi}_i \cdot W_{T,i}^{\gamma}}{e^{-rT} \sum_i \bar{\pi}_i \cdot W_{T,i}^{\gamma}} = \frac{\bar{\pi}_i \cdot W_{T,i}^{\gamma}}{\sum_i \bar{\pi}_i \cdot W_{T,i}^{\gamma}} \tag{9.7}$$

While $\bar{\pi}_i$ can be extracted from observable option prices, it is much more difficult to come up with an estimate for γ. It is well known from empirical work on the equity risk premium puzzle[8] that the smoothness of consumption results in unreasonable high values for γ. In other words, as consumptions does vary little between states of the world, investors need to be very risk averse to justify the high historical risk premium. We take a more pragmatic approach instead. Varying the risk aversion γ in Equation 9.7 allows us to calibrate the expected return under the real-world distribution to the targeted

Table 9.2 From risk-neutral to real-world distribution

$W_{T,i}$	$\bar{\pi}_i$ (%)	$\bar{\pi}_i W^g_{T,i}$	π_i (%)
⋮			
4,875.0	5.35	12,70,916.57	3.13
5,125.0	6.55	17,19,943.89	4.24
5,375.0	7.50	21,65,634.19	5.34
5,625.0	8.09	25,60,099.25	6.31
5,875.0	8.30	28,63,247.90	7.05
6,125.0	8.13	30,49,525.06	7.51
6,375.0	7.65	31,10,385.49	7.66
6,625.0	6.96	30,52,867.93	7.52
6,875.0	6.13	28,95,589.33	7.13
7,125.0	5.25	26,63,723.96	6.56
7,375.0	4.38	23,84,261.88	5.87
7,625.0	3.58	20,82,337.72	5.13
7,875.0	2.87	17,78,919.21	4.38
8,125.0	2.26	14,89,773.70	3.67
8,375.0	1.75	12,25,427.05	3.02
8,625.0	1.33	9,91,765.45	2.44
⋮			

Real-world probabilities for $\gamma = 2$ are calculated using Equation 9.7 with the same parameters as in Table 9.1.

level

$$\left(\frac{1}{S_0}\right)\sum_i \pi_i \frac{S_i + S_{i+1}}{2} - 1$$

For $\gamma = 2$ we arrive at an expected return of 12.77%. As mentioned before, we use γ simply as a calibration parameter.

9.4 NUMERICAL APPLICATION

We suggest solving the following problem. Minimise the tracking error between the continuous CPPI trading strategy and our static buy and hold tracking portfolio

$$\min \sum_i \pi_i (W^*_{T,i} - w_{\mathrm{C}} \mathrm{e}^{rT} - w_{\mathrm{S}} S_{T,i} - w_{\mathrm{C}_1} C_{1,i} \cdots)^2 \tag{9.8}$$

subject to a budget constraint

$$W_0 = \mathrm{e}^{-rT} \sum_i \bar{\pi}_i (w_{\mathrm{C}} \cdot 6{,}000\mathrm{e}^{rT} + w_{\mathrm{S}} S_{T,i} + w_{\mathrm{C}_{1,i}} \cdots) = 6{,}000 \tag{9.9}$$

and a non-negativity constraint on individual asset holdings

$$w_{\mathrm{c}} \geqslant 0, \qquad w_{\mathrm{s}} \geqslant 0, \qquad w_{\mathrm{c}_1} \geqslant 0, \qquad \dots, \qquad w_{\mathrm{c}_m} \geqslant 0 \tag{9.10}$$

as well as a cardinality constraint for the number of instruments (n_{assets}) involved

$$\left.\begin{aligned} w_{\mathrm{c}} &\leqslant \text{large number} \cdot \delta_{\mathrm{c}}, \quad \delta_{\mathrm{c}} \in \{0,1\} \\ w_{\mathrm{s}} &\leqslant \text{large number} \cdot \delta_{\mathrm{s}}, \quad \delta_{\mathrm{s}} \in \{0,1\} \\ w_{\mathrm{c}_1} &\leqslant \text{large number} \cdot \delta_{\mathrm{c}_1}, \quad \delta_{\mathrm{c}_1} \in \{0,1\} \\ &\vdots \\ w_{\mathrm{c}_m} &\leqslant \text{large number} \cdot \delta_{\mathrm{c}_m}, \quad \delta_{\mathrm{c}_m} \in \{0,1\} \end{aligned}\right\} \tag{9.11}$$

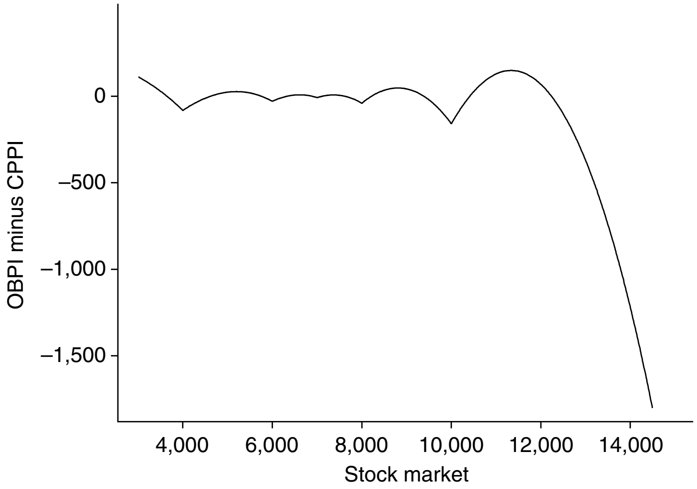
Note that each expression in Equation 9.11 provides a logical switch using a binary variable that takes on a value of either 0 or 1. Let us review the case for cash to clarify the calculations. As soon as $w_{\mathrm{c}} > 0$ by even the smallest amount, cash enters the optimal solution in which case $\delta_{\mathrm{c}} = 1$ to satisfy the inequality. If $w_{\mathrm{c}} = 0$ instead, it must follow that $\delta_{\mathrm{c}} = 0$. Computationally the "large number" should not be chosen "too large", ie, it depends on how large w_{c} can become. Finally we need to add the "dummy variables" to count the number of assets that enter the optimal solution

$$\delta_{\mathrm{c}} + \delta_{\mathrm{s}} + \delta_{\mathrm{c}_1} + \cdots + \delta_{\mathrm{c}_m} \leqslant n_{\text{assets}} \tag{9.12}$$

After having specified the optimisation problem (Equations 9.8–9.12) we can solve our tracking problem for various numbers of admissible instruments to see how many instruments we need to track a given CPPI profile with a static OBPI (options-based portfolio insurance) portfolio. We use the same numerical example as in the second section. Strikes for the respective call options are assumed to range between 2,000 and 12,000 with a step size of 1,000 points (m = 11 options).

In Figure 9.2 we can observe the logic of Equation 9.8 in the case of $n_{\text{assets}} = 6$. Tracking optimisation does not mean to minimise payout differences across the whole spectrum of stock market outcomes but rather it focuses on the more likely outcomes. Note that hedging errors (difference between CPPI and OBPI payouts) are small around the current stock price of 6,000 and much larger where tracking is less relevant, ie, where the real-world probability is low. In the case when traded option prices exhibit a skewed distribution, we would have

Figure 9.2 Hedging error of a static option hedge

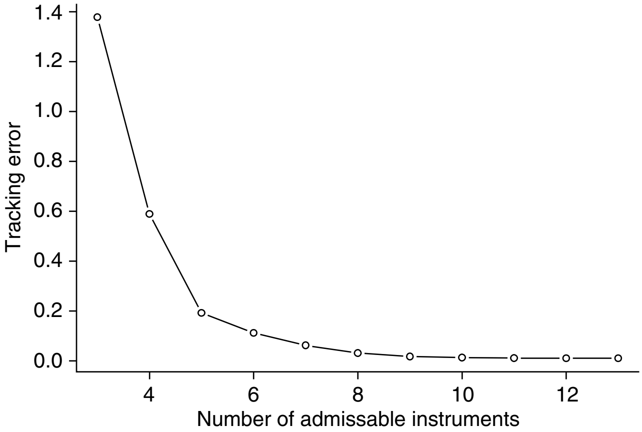


We used $n_{\text{assets}} = 6$ to track a given CPPI strategy. Note that hedging errors (difference between CPPI and OBPI payouts) are small around the current stock price of 6,000 and much larger where tracking is less relevant, ie, where the real-world probability is low.

to rethink our optimal static hedge some outcomes might become more important.[9]

Repeating this exercise for various numbers of admissible assets (ranging from 2 to 13) leads us to Figure 9.3. As the number of instruments rises, the tracking error decreases. Note that tracking error is expressed as annual percentage volatility, meaning 0.2 equals 20%. The tracking advantage tails off quickly with more than eight admissible instruments. Exact numbers are given in Table 9.3. The first column describes the available instruments. For example $C_{5,000}$ denotes the call option with a strike of 5,000. For each number of available instruments (stated in the top row) the associated positions are quoted. Allowing three instruments will result in a cash position as long as two long calls to mimic the convexity of a CPPI strategy. Given that only two instruments are allowed the optimisation uses two widely separate calls to mimic the convex payout of a CPPI strategy. For this special case our optimisation suggests to invest $0.85 \cdot 6{,}000 = 5{,}100$ into cash plus a long call position of 0.665 units of an in-the-money call with strike 5,000 and another long call position of 0.784 units of an out-of-the-money call with strike 8,000.

Figure 9.3 Reduction of hedging error and number of admissible instruments



As the number of instruments rises, the tracking error decreases. Note that the tracking error is expressed as annual percentage volatility, meaning 0.2 equals 20%. The tracking advantage tails off quickly with more than eight admissible instruments.

Let us check the budget constraint by calculating both call prices with the help of Table 9.1

$$\begin{aligned} E_Q(\max(S_{T,i} - 5{,}000.0)) &= \mathrm{e}^{-rT} \sum_i \bar{\pi}_i \max(S_{T,i} - 5{,}000.0) \\ &= 1{,}268.94 \\ E_Q(\max(S_{T,i} - 8{,}000.0)) &= \mathrm{e}^{-rT} \sum_i \bar{\pi}_i \max(S_{T,i} - 8{,}000.0) \\ &= 70.60 \end{aligned}$$

Adding up the value of these three investments must yield our initial wealth of 6,000. Otherwise we used too much capital to generate a tracking portfolio

$$0.85 \cdot 6{,}000 + 0.655 \cdot 1{,}268.94 + 0.784 \cdot 70.60 = 6{,}000$$

If we allow an additional investment moving to the right in Table 9.3, our optimisation procedure will increase the number of investments. Convexity is first added around the most likely scenarios, ie, around the current stock price. Not also that at no time a direct stock investment is made, as it simply offers no convexity. The amount of cash necessary to preserve a minimum wealth decreases with the number of purchased call options.

Table 9.3 Optimal static hedges

	Number of instruments									
	3	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
$C_{3,000}$	0.000	0.000	0.000	0.000	0.000	0.304	0.304	0.304	0.192	0.192
$C_{4,000}$	0.000	0.524	0.406	0.426	0.427	0.000	0.000	0.000	0.128	0.128
$C_{5,000}$	0.665	0.000	0.000	0.000	0.000	0.176	0.176	0.176	0.157	0.157
$C_{6,000}$	0.000	0.000	0.357	0.263	0.257	0.184	0.185	0.185	0.189	0.189
$C_{7,000}$	0.000	0.482	0.000	0.167	0.200	0.227	0.221	0.222	0.221	0.221
$C_{8,000}$	0.784	0.000	0.500	0.368	0.235	0.226	0.256	0.252	0.253	0.253
$C_{9,000}$	0.000	0.680	0.000	0.000	0.397	0.400	0.258	0.284	0.284	0.284
$C_{10,000}$	0.000	0.000	0.745	0.804	0.000	0.000	0.441	0.320	0.320	0.316
$C_{11,000}$	0.000	0.000	0.000	0.000	0.858	0.857	0.000	0.312	0.312	0.348
$C_{12,000}$	0.000	0.000	0.000	0.000	0.000	0.000	0.893	0.559	0.559	0.370
$C_{13,000}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.471
S_0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Cash	0.850	0.791	0.812	0.809	0.809	0.774	0.774	0.774	0.790	0.790

The first column describes the available instruments. For example $C_{5,000}$ denotes the call option with a strike of 5,000. For each number of available instruments (stated in the top row) the associated positions are quoted. Allowing three instruments will result in a cash position of $0.85 \cdot 6{,}000 = 5{,}100$, a long call position of 0.665 units of an in-the-money call with strike 5,000 and another long call position of 0.784 units of an out-of-the-money call with strike 8,000.

9.5 EXTENSIONS TO UTILITY OPTIMISATION

The previous section assumed log-normality for the risk-neutral density rather than deriving it from market data. Option prices on the other side provide timely information on the markets implied risk-neutral density for the expiry date. The literature on portfolio optimisation uses exclusively historical data to estimate the distribution of asset returns imposing distributional assumptions on the marginal distribution of asset returns. This ignores the markets expectations about both the level of volatility as well as the shape of this distribution, which might not only deviate considerably from log-normality, but might also vary considerably across time, so that simply fitting a non-normal alternative to the historical data will not be able to catch the dynamics of changing expectations. We suggest therefore merging the option pricing literature which has long dealt with deviations from log-normality when building implied trees[10] with portfolio optimisation algorithms to eliminate this deficiency. The following section will use the implied real-world distribution from option prices to find the optimal equity/cash split for an investor with CRRA utility across different risk aversions. We will contrast our results with the case of a lognormal utility.

While many methods to extract risk-neutral densities exist, we focus on the most common parametric method based on Shimko (1993). We know that implied volatilities deviate from being constant when the risk-neutral density deviates from log-normality. Hence, we first fit a functional form to implied volatility and then back out the corresponding risk-neutral density. For the relationship between implied volatility and strike price we use

$$\hat{\sigma}_{\text{impl}}(X) = \hat{a}_1 + \hat{a}_2 X + \hat{a}_3 X^2 \tag{9.13}$$

where $\hat{a}_1, \hat{a}_2, \hat{a}_3$ are the estimated parameters from a linear (cross-sectional) regression of the implied volatility σ_{impl} against strike and strike squared. In the second step the risk-neutral probability can then be approximated by[11]

$$\bar{\pi} = \mathrm{e}^{rT}\Bigg(\frac{C(S_{i+1}, \hat{\sigma}_{\text{impl}}(S_{i+1}), \dots)}{(S_{i+1} - S_i)^2} - \frac{2C(S_i, \hat{\sigma}_{\text{impl}}(S_i), \dots)}{(S_{i+1} - S_i)^2} + \frac{C(S_{i-1}, \hat{\sigma}_{\text{impl}}(S_{i-1}), \dots)}{(S_{i+1} - S_i)^2}\Bigg) \tag{9.14}$$

Table 9.4 Sample call option data

Strike	σ_{impl}
5,125	0.2551
5,325	0.2181
5,525	0.1632
5,625	0.1701
5,725	0.0968
5,825	0.1478
5,925	0.1349
6,025	0.1272
6,125	0.1047
6,225	0.1121
6,325	0.1061
6,425	0.1028
6,525	0.0991
6,625	0.0961
6,725	0.0968
6,825	0.1006

The table shows strike price and implied volatility for call options of different strikes. This data is used in the cross-sectional regression (Equation 9.13). Note that $T = (1.5/1.2)$, ie, option maturity is assumed to be 1.5 months.

Table 9.5 Quadratic regression to fit implied volatility

	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	R^2
OLS	3.61	−10.85	8.36	0.90
Robust regression	3.32	−9.77	7.39	0.95

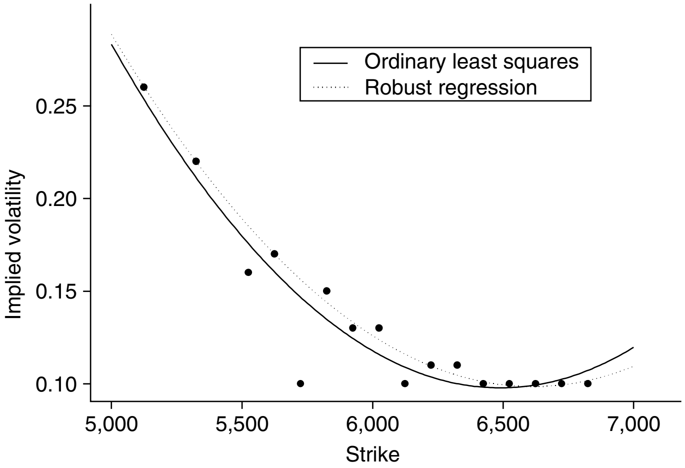
We show estimates of Equation 9.13 for the data in Table 9.4.

where $C(S_i, \hat{\sigma}_{\text{impl}}(S_i), \dots)$ denotes the price of European call with strike price S_i, the corresponding fitted implied volatility $\hat{\sigma}_{\text{impl}}(S_i)$ according to Equation 9.13 and all other parameters like maturity, risk-free rate, dividend yield and current stock price, S_0, being the same as in the second section. Suppose we are given a set of implied volatilities across a series of strikes as in Table 9.4.

In Step 1 we perform both robust as well as simple OLS regressions.[12] The coefficient estimates are given in Table 9.5.

Studying the results of both regressions in Figure 9.4, we decide to use the robust regression as the effect of outliers seem to substantially

Figure 9.4 Fitted implied volatilities for linear and robust regression



The cross-sectional regression (Equation 9.13) is run using both OLS and robust regression. Fitted lines are either dashed or dotted, while the raw data is provided in Table 9.4.

distort the results. In Step 2 we use Equation 9.14 in combination with Equation 9.13 to estimate risk-neutral probabilities. For means of comparison we also plot the risk-neutral probability build on the assumption of log normality and at-the-money volatility (annual volatility of 11%).

We can see that the option implied risk-neutral probability is more heavily skewed given a larger weight to negative returns than its lognormal alternative. For each distribution a different γ is chosen in order to equate the expected return of both distributions. Otherwise any comparison would be spurious. Note that we cannot use the risk-neutral distribution for utility optimisation. Any risk-averse investor would invest 100% in the riskless asset as the return expectation for all assets equals the risk-free rate.

Finally (step 3) we compute the utility maximising solution for an investor with CRRA utility $u = (1/(1-\upsilon))W^{1-\upsilon}$ for $\upsilon \neq 1$. Our problem becomes to find the optimal equity/cash allocation for realistic return distributions, ie

$$\max \sum_i \pi_i \frac{1}{1-\upsilon}(W_{T,i})^{1-\upsilon} \tag{9.15}$$

Table 9.6 Optimal equity allocations

Risk aversion (ν)	Implied volatility (%)	Lognormal volatility (%)
2	100.00	100.00
5	74.54	78.91
10	37.51	39.45
50	7.54	7.89

Allocations are given by the solution of Equations 9.15 and 9.16 for different levels of risk aversions. Real-world scenarios are calibrated using Equation 9.7 for both the option implied and a standard lognormal risk-neutral distribution.

subject to our well-known budget (valuation) constraint[13]

$$W_0 = \mathrm{e}^{-rT} \sum_i \bar{\pi}_i (w_{\mathrm{C}} \cdot 6{,}000\mathrm{e}^{rT} + w_{\mathrm{S}} S_{T,i}) = 6{,}000 \tag{9.16}$$

that ensures that we do not overspend on assets, that all assets are priced according to market conditions and that real-world probabilities are consistent with risk-neutral probabilities. It should be clear that the non-normality associated with the option implied real-world distribution will command a different solution. The larger possibility of extreme losses implied by the option data in Table 9.4 are likely to call for a less aggressive equity allocation.

The results in Table 9.6 confirm our intuition. The implied distribution with its skew towards negative returns leads to less aggressive allocations than the standard assumption of log-normality. Even in the case of extremely low risk aversion, it is essentially the non-negativity constraint that equates both solutions. If shorts were admissible 196% would be invested in equities under log-normality, while the option implied distribution would favour 182% in equities.

9.6 SUMMARY

This chapter provides a straightforward methodology of relating CPPI strategies to OBPI. Without having to rely on stochastic simulations we can evaluate how close a static hedge using a limited number of options will come to tracking a hypothetical CPPI in the most favourable circumstances for the CPPI. The advantage of this approach is that as soon as the replicating portfolio is set up we are agnostic about transaction costs or price jumps. As a side effect, this methodology offers an intuitive way to merge option pricing and portfolio optimisation technology. In our opinion this is the correct way to deal with non-normality. A scenario based utility optimisation, where scenarios are derived from fitted risk-neutral distribution. Fitting marginal distributions to historical data imposing distributional assumptions and using questionable risk measures to account for non-normality will hopefully be a method of the past or confined to those assets where liquid option markets do not exist.

EXERCISE

1. Take the data in Table 9.4. and find the option portfolio that maximises the expected long-term portfolio growth (log utility) under the assumed real-world distribution.

[1] See, for example, Bookstaber and Langsam (2000) or Black and Perold (1992).

[2] Lo and Haugh (2001) also use options to track the payout of optimal dynamic investment policies for alternative asset processes and utility functions. In contrast to Lo and Haugh we use mixed integer programming with cardinality constraints, rather than searching for the optimal solution by solving all possible sub-problems. Also we are more explicit about creating consistent scenarios for both the risk-neutral as well as the real-world probability measure. Finally we suggest merging option pricing theory and the portfolio optimisation literature to incorporate market expectations about extreme returns, while still pricing all assets correctly.

[3] See Bertrand and Prigent (2002).

[4] Scherer and Martin (2005) provide sample solutions for this class of problems using NuOPT for S-Plus. All calculations are performed with NuOPT for S-Plus. Code is available from the author on request.

[5] See Jensen (1972, p. 357).

[6] See Zhang (1997, p. 391f).

[7] See Zimmermann (1998, p. 76). Note that this actually assumes that

$$u(W) = \frac{1}{1-\gamma} W^{1-\gamma} \quad \text{for } \gamma \neq 1$$

which results in $\mathrm{d}u/\mathrm{d}W = w^{-\gamma}$.

[8] See Mehra and Prescott (1985).

[9] Under these circumstances Equation 9.1 would be unsuitable to calculate the CPPI payout across different scenarios. Instead we would be advised to fit an implied tree to observable option prices, and calculate the CPPI strategies as we move forward through the tree.

[10] See Jackwerth (1999) for a comprehensive review.

[11] See Taylor (2005, p. 446).

[12] See Rousseeuw and Leroy (1987) for robust regression.

[13] We also include a non-negativity constraint on cash and equity investments to reflect a realistic investor setting.

REFERENCES

Bertrand, P., and J. Prigent, 2002, "Portfolio Insurance Strategies: OBPI versus CPPI", Discussion Paper, GREQAM and Université Montpellier.

Black, F., and A. Perold, 1992, "Theory of Constant Proportion Portfolio Insurance", *Journal of Economic Dynamics and Control* 16, pp. 403–26.

Bookstaber, R., and J. Langsam, 2000, "Portfolio Insurance Trading Rules", *Journal of Futures Markets* 8, pp. 15–31.

Jackwerth, J., 1999, "Option-Implied Risk-Neutral Distributions and Implied Binomial Trees: Literature Review", *Journal of Derivatives* 7, pp. 66–82.

Jensen, P., 1972, "Capital Markets, Theory and Evidence", *Bell Journal of Economics and Management Science* 3, pp. 357–98.

Lo, A., and M. Haugh, 2001, "Asset Allocation and Derivatives", *Quantitative Finance* 1, pp. 45–72.

Mehra, R., and E. Prescott, 1985, "The Equity Premium Puzzle", *Journal of Monetary Economics* 15, pp. 145–61.

Rousseeuw, P., and A. Leroy, 1987, *Robust Regression and Outlier Detection* (New York: John Wiley & Sons).

Scherer, B., and D. Martin, 2005, *Introduction to Modern Portfolio Optimization with NuOPT™, S-Plus®, and S^+ Bayes™* (New York: Springer).

Shimko, D., 1993, "Bounds of Probability", *Risk* 6, pp. 33–7.

Taylor, S., 2005, *Asset Price Dynamics, Volatility, and Prediction* (Princeton, NJ: Princeton University Press).

Zhang, P., 1997, *Exotic Options* (Hackensack, NJ: World Scientific).

Zimmermann, H., 1998, *State Preference Theory and Asset Pricing* (Berlin: Physica).

10
Scenario Optimisation

Scenario optimisation is used where the parameters of a mathematical model (asset returns over the next year, for example) are subject to randomness. One way to solve these stochastic programmes is to solve a deterministic problem with many different scenarios assumed for the uncertain inputs – in our case, returns. We can, for example, sample 100,000 scenarios for five assets from the predictive distribution (see Chapter 6) to arrive at a $100{,}000 \times 5$ cell matrix. After the draws have been made, uncertainty is removed and we are left with the deterministic problem of which asset weights to choose in order to maximise the objective, taking into account what can happen in all 100,000 scenarios. We will see later that for many objectives this can be written as a linear program.[1]

Scenario optimisation investigates how well feasible solutions perform in different scenarios (using an objective function to evaluate the solution for each scenario) and then tries to find the optimal solution. Although future asset returns are uncertain, scenario optimisation attempts to draw representative scenarios that could happen in the future. Scenario optimisation that is based on only a few scenarios may overadjust to the data, providing too optimistic an assessment of what can be achieved, while scenario optimisation with a very large set of scenarios can be computationally infeasible.[2]

The first section focuses on utility-based models, while the second section reviews basic issues in scenario generation. The remainder of the chapter focuses on risk perceptions that are widely used in asset management applications and can only be tackled using scenario optimisation.[3]

10.1 UTILITY-BASED SCENARIO OPTIMISATION

Utility-based models have a strong foundation in decision theory as they directly maximise expected utility rather than relying on approximations (see Chapter 1). Expected utility allows us to deal with arbitrarily shaped distributions rather than forcing us to deal with mean and variance only.[4]

We define expected utility, EU, as

$$\begin{aligned} \mathrm{EU}(W) &= \sum_{i=s}^{S} p_s U(1 + R_{\mathrm{ps}}) = \frac{1}{S} \sum_{i=s}^{S} U(1 + R_{\mathrm{ps}}) \\ &= \frac{1}{S} \sum_{i=s}^{S} U\left(1 + \sum_{i=1}^{k} w_i R_{is}\right) \end{aligned} \tag{10.1}$$

where W denotes end-of-period wealth and R is the return of asset i in scenario s. Equation 10.1 states that expected utility is calculated for $s = 1, \dots, S$ scenarios. Each scenario is drawn with probability $p_s = 1/S$ and gives a portfolio return of R_{ps}. Utility is specified as a utility function with constant relative risk aversion

$$U(1 + R) = \begin{cases} \dfrac{(1 + R)^{1-\gamma}}{1 - \gamma}, & \gamma \geqslant 0 \\ \ln(1 + R), & \gamma = 1 \end{cases} \tag{10.2}$$

where γ is the risk-aversion coefficient. Any series of historic (or simulated) excess returns can be rewritten as

$$R_{is} = c + \mu_i + \sigma_i z_{is} \tag{10.3}$$

where c is the return of a risk-free alternative, μ_i is the expected unconditional excess return of the ith of $i = 1, \dots, k$ assets and σ_i is the ith asset's volatility. With z_{is} we describe the standardised random part of the excess returns, ie, we take the original series of excess returns, subtract its mean and divide the result by its volatility. Thereby we create a series with a mean of zero and a volatility of one that retains all non-normality in the data – standardisation does not change the shape of a distribution. However, Equation 10.3 helps us to control the first two moments, ie, we can change the location (mean) and the dispersion parameter (volatility) as we like – for example, to allow for expectations of lower risk premiums in the future – without changing the degree of non-normality in the data. If we wished we could also input a new risk-free rate, a new expected return or a new forward-looking volatility figure.

A similar decomposition is readily available for benchmark (portfolio) returns, $R_{\mathrm{b}s}$

$$R_{\mathrm{b}s} = c + \mu_{\mathrm{b}} + \sum_{i=1}^{k} w_i \sigma_i z_{is} \tag{10.4}$$

where μ_{b} is the expected risk premium on a given portfolio and $\sum_{i=1}^{k} w_i \sigma_i z_{is}$ describes the innovations (deviations from the expected benchmark return) as a sum of shocks on all k assets. If we regard these benchmark portfolio positions as optimal (zero alpha) – ie, we assume that Equation 10.1 has been optimised – we can find a valuation model from Equation 10.1 that equalises all expected returns under a set of modified probabilities, p_s^*[5]

$$P_s^* = \frac{p_s(\mathrm{d}U/\mathrm{d}(1+R_{\mathrm{b}s}))}{\sum_{i=1}^{S} p_s(\mathrm{d}U/\mathrm{d}(1+R_{\mathrm{b}s}))} = \frac{p_s(1+R_{\mathrm{b}s})^{-\gamma}}{\sum_{i=1}^{S} p_s(1+R_{\mathrm{b}s})^{-\gamma}} \tag{10.5}$$

so that

$$\sum_{s=1}^{S} p_s^* R_{is} = \sum_{s=1}^{S} p_s^* R_{\mathrm{b}s} \tag{10.6}$$

Substituting Equations 10.3 and 10.4 into Equation 10.6, we arrive at the utility-based version of "grapes from wine", ie, we arrive at the implied returns a scenario-based utility maximiser must have used to find the current portfolio weights optimal[6]

$$\sum_{s=1}^{S} p_s^*(c + \mu_i + \sigma_i z_{is}) = \sum_{s=1}^{S} p_s^*\left(c + \mu_{\mathrm{b}} + \sum_{i=1}^{k} w_i \sigma_i z_{is}\right)$$

Recalling that $\sum_{s=1}^{S} p_s^* = 1$ we can solve for μ_i to get

$$\mu_i = \mu_{\mathrm{b}} + \sum_{s=1}^{S} p_s^*\left(\sum_{i=1}^{k} w_i \sigma_i z_{is}\right) - \sum_{s=1}^{S} p_s^*(\sigma_i z_{is}) \tag{10.7}$$

for all $i = 1, \dots, k$. Changing the risk-aversion parameter will change marginal utility and, hence, p_s^* and the required implied returns.

Let us illustrate the procedure with the data tabulated in Appendix A (on page 298). To obtain enough data points to span the sample space meaningfully, we bootstrap 500 annual excess returns from the raw data. This exercise yields the distributional characteristics for the sampled excess returns summarised in Table 10.1.[7] We also assume benchmark weights of 40% for the UK, 30% for Japan, 10% for Europe and 20% for the US.

Table 10.1 Distributional characteristics of sampled excess returns for monthly return series given in Appendix A (on page 298)

Correlation

	UK	**Japan**	**Europe**	**US**
UK	1.00			
Japan	0.22	1.00		
Europe	0.72	0.32	1.00	
US	0.59	0.26	0.58	1.00

Distribution

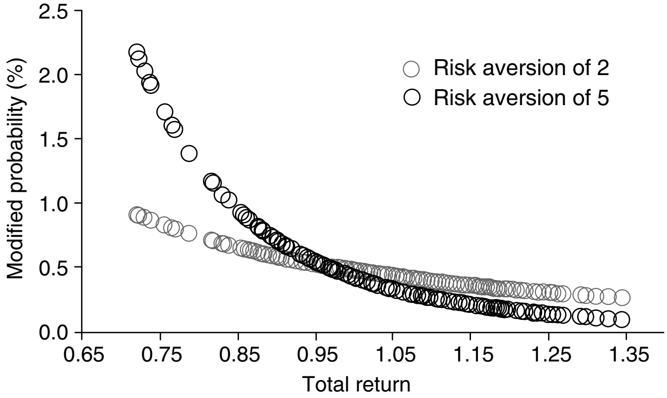
	UK	**Japan**	**Europe**	**US**
Volatility (%)	14.99	19.58	17.51	15.46
Skewness	0.11	0.79	0.27	0.12
Kurtosis	0.29	1.86	−0.16	−0.16

Next we want to calculate the modified probabilities for different risk aversions as described in Equation 10.5. Figure 10.1 presents the results for two different risk aversions.

Higher risk aversions lead to a significant increase in modified probability for negative market returns (remember that Figure 10.1 plots $1 + R$ rather than R alone). The more risk averse an investor is, the more weight (higher adjusted probability) will be given to the left tail of the distribution. If assets show non-normality, the returns required to accept a given portfolio as optimal will change. This, in turn, will lead to changed implied returns, as calculated in the second and third terms of Equation 10.7. While the second term looks at the required risk premium that can be attributed to the benchmark return distribution, the third term catches the effect that can be attributed to the ith asset itself. Implied returns (returns that are required to compensate for risk aversion and, in this case, also non-normality) will rise if, on average, deviations from expected returns, z_{is}, are highly negative when p_s^* is also high. Recall that p_s^* is high if marginal utility is high and, hence, deviations from expected benchmark returns, z_{b}, are low. This will be the case if z_i and z_{b} are positively correlated (or, in some states, negatively).

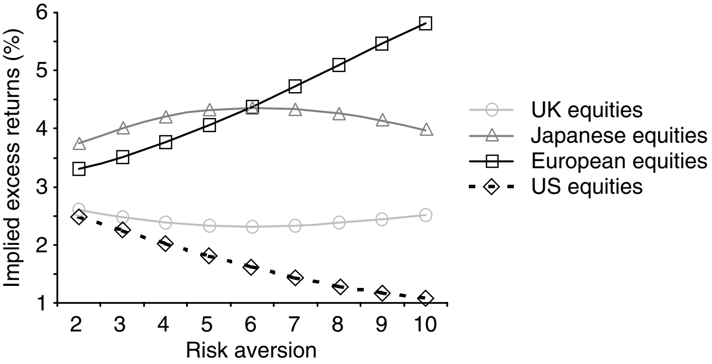
We can now look directly at how implied returns (obtained using Equation 10.2) change as we alter risk aversions. The results are

Figure 10.1 Modified probabilities for different risk aversions



This and subsequent figures are based on monthly return series given in Table 10.5 of Appendix A (on page 298).

Figure 10.2 Implied returns for alternative risk aversions



summarised in Figure 10.2. We assume a risk premium of 3% for the optimal portfolio together with a risk-free rate of 3.5%.

As risk aversion rises, the required returns for the most risky assets (Japanese and European equities) also increase, though not in a linear fashion (as would be the case for ordinary implied excess returns if we used a Markowitz framework). For example, we see that the implied excess return for UK equities – the asset with the lowest volatility – falls for intermediate risk aversions but that, as risk aversion increases further, it rises again. The reason for this is that UK equities are also the asset with the least positive skewness. For rising risk aversions we have to expect considerable return differences to justify the current portfolio holdings.

In this section we have seen that implied returns can be backed out from actual allocations even in a utility-based scenario optimisation framework. However, as most investors feel somewhat uncomfortable specifying their utility functions (most readers will find it difficult to find a value of γ that suits them in Equation 10.2), we will move away from utility-based to risk-based models for the remainder of this chapter.

10.2 SCENARIO GENERATION

Scenario optimisation techniques turn a stochastic problem into a deterministic problem by simulating future scenarios for all assets.[8] The problem becomes deterministic as soon as the simulation stops and all scenarios have been generated. In this form the problem can now be solved using mathematical (mostly linear) programming techniques. Key to the quality of the solution we get is the quality of the sampled scenarios. In particular, scenarios must be

- parsimonious – one should use a relatively small number of scenarios to save computing time;
- representative – scenarios must offer a realistic description of the relevant problem and not induce estimation error; and
- arbitrage-free: scenarios should not allow the optimiser to find highly attractive solutions that make no economic sense.

Before we look at scenario optimisation techniques we will therefore concentrate for a few pages on scenario generation.

Scenario generation can draw on an array of proven methods. First, we could use Bayesian methods as described in Chapter 6 to model the predictive distribution. Theoretically, the predictive distribution is ideal as it is exactly this distribution – which incorporates uncertainty about estimated parameters as well as prior knowledge about the future – that defines expected utility. In practice, however, investors might find Bayesian methods technically challenging, and priors always leave the impression of arbitrariness and are difficult to formulate, particularly as one also always needs to specify the degree of belief. Furthermore, the predictive distributions of Chapter 6 will be close to a normal distribution and so might not reflect the level of non-normality in the data.

A second method is to use bootstrapping to generate future scenarios, calculating annual returns from 12 bootstrapped monthly returns and repeating this procedure 1,000 times to get a distribution of future annual returns.[9] Although bootstrapping leaves correlations unchanged, it will, by construction (as it imposes the independence assumption of the Central Limit Theorem on the resampled data), destroy all dependencies between two return realisations, such as autocorrelation in returns, Garch effects, etc. Recall that bootstrapping is done by repeated independent draws, with replacement.[10] The data will therefore look increasingly normal after a bootstrapping exercise even if they were not normally distributed in the first place.

Instead of bootstrapping from the empirical distribution, we could use Monte Carlo simulation, ie, draw from a parametric distribution. This can be done by fitting a distribution to each asset under consideration and then generating a multivariate distribution by sampling from the individual distributions separately, while gluing distributions together according to a correlation matrix using rank correlations.

Alternatively, we could estimate a vector–autoregressive model of the form[11]

$$\boldsymbol{r}_t = \boldsymbol{\mu} + \boldsymbol{\Theta}(\boldsymbol{\mu} - \boldsymbol{r}_{t-1}) + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim N(\boldsymbol{0}, \boldsymbol{\Omega}) \tag{10.8}$$

where $\boldsymbol{\Theta}$ is the matrix of autoregressive parameters, $\boldsymbol{\mu}$ defines mean returns, $\boldsymbol{r}_{t-1}$ defines lagged returns and $\boldsymbol{\Omega}$ describes the covariance structure of the error terms, $\boldsymbol{\varepsilon}$. While Equation 10.8 is mainly used to develop multiperiod scenarios for multiperiod optimisation (not discussed here but standard in the asset–liability modelling literature) because it introduces path-dependency into asset returns, it also proves useful in one-period models where significant path-dependence has considerable influence on the riskiness of different assets even over one period.[12] It has been shown that as long as the covariance matrix from which we sample is non-singular, the conditional variance of any asset with respect to all other assets is positive.[13] This is the same as saying that no asset can be perfectly explained (replicated) by a combination of the other assets, and it therefore rules out arbitrage.

10.2.1 Eliminating arbitrage

An important consideration in generating scenarios is to reduce or eliminate the possibility of arbitrage. Suppose we used one of the above methods to generate $s = 1, \ldots, S$ scenarios for k assets. All scenarios are summarised in a scenario matrix, $\boldsymbol{S}_{S\times k}$, that contains $S \times k$ asset returns

$$\boldsymbol{S} = \begin{bmatrix} 1+c+r_{11} & 1+c+r_{12} & \cdots & 1+c+r_{1k} \\ 1+c+r_{21} & 1+c+r_{22} & \cdots & 1+c+r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1+c+r_{S1} & 1+c+r_{S1} & \cdots & 1+c+r_{Sk} \end{bmatrix} \tag{10.9}$$

For notational convenience in later equations we let

$$\boldsymbol{R}_s' = \begin{pmatrix} c+r_{s1} & \cdots & c+r_{sk} \end{pmatrix}$$

denote the vector of returns across all k assets in scenario s. Although arbitrage is ruled out if we sample from a non-singular covariance matrix, this is strictly true only for a large number of drawings as sampling error for a small number of drawings might result in arbitrage. To test for arbitrage opportunities – which the optimiser will certainly find and will leverage on – we have first to understand what arbitrage means in the context of our scenario matrix, $\boldsymbol{S}_{S\times k}$. Suppose that we live in a world with two assets and two scenarios. $\boldsymbol{S}^1$ and $\boldsymbol{S}^2$, below, are examples of two scenario matrixes

$$\boldsymbol{S}^1 = \begin{bmatrix} 1+10\% & 1+7\% \\ 1+6\% & 1+5\% \end{bmatrix} \quad \text{and} \quad \boldsymbol{S}^2 = \begin{bmatrix} 1+10\% & 1+8\% \\ 1+4\% & 1+5\% \end{bmatrix}$$

Both scenarios in each matrix are equally likely. An arbitrage opportunity can be loosely defined as a portfolio that is guaranteed not to lose money at the end of the investment period in any scenario but which generates a positive cashflow at the beginning (the investor receives money by cleverly setting up that arbitrage portfolio by selling and buying both assets).[14] To find arbitrage mathematically, we solve

$$\min_{W} \boldsymbol{w1}, \qquad \boldsymbol{Sw} \geqslant \boldsymbol{0} \tag{10.10}$$

where $\boldsymbol{1}$ is a vector of 1s. The minimisation objective in Equation 10.10 expresses the prize of an arbitrage portfolio. All assets are assumed to cost one monetary unit. If the objective is negative, the investor receives money. The constraints apply to the payouts in both scenarios. Solving Equation 10.10 for both scenario matrixes we see that the first scenario matrix, $\boldsymbol{S}^1$, incorporates an arbitrage opportunity in that the optimiser will not converge, while the second, $\boldsymbol{S}^2$, does not. This is not surprising, as it is clear that in scenario matrix $\boldsymbol{S}^1$ trivially asset one dominates asset two in both scenarios, giving higher returns in both states of the world; hence, long asset one and short asset two will do the trick. However, as the number of assets and scenarios rises, situations like this become more difficult to pick out, so we will settle for Equation 10.10 to test for arbitrage opportunities.[15] The likelihood of encountering arbitrage opportunities increases as the number of scenarios decreases, but this problem can be eliminated by sampling additional states of the world until arbitrage opportunities disappear. Basically, adding additional columns to the scenario matrix will make it unlikely that we will find a combination of assets that allows arbitrage. Also, limiting ourselves to long-only portfolios – as they apply to most institutional investors – will avoid unbounded solutions (in the presence of arbitrage the arbitrage profit becomes infinite) but will still result in solutions where some assets are overrepresented.

10.3 MEAN–ABSOLUTE DEVIATION

The first scenario-based alternative to Markowitz optimisation we will consider is the mean–absolute deviation (MAD) model.[16] This measures risk as an absolute deviation from the mean rather than squared deviation as in the case of variance. We can define MAD as

$$\text{MAD} = \sum_{s=1}^{S} p_s |\boldsymbol{w}'[\boldsymbol{R}_s - \bar{\boldsymbol{R}}]| \tag{10.11}$$

where p_s denotes the probability of scenario s, $\boldsymbol{w}$ is the vector of portfolio weights (in all scenarios), $\boldsymbol{R}_s$ is the vector of k asset returns in scenario s and $\bar{\boldsymbol{R}}$ is the vector of asset means across all scenarios. As we work with simulated scenarios with equal probability, we will always set $p_s = 1/S$. Whereas in variance-based methods the property of the square penalises larger deviations at an increasing rate, this is not the case with MAD. In fact MAD implies that a further unit of underperformance relative to the mean creates the same disutility no matter how big the loss already is. However, one advantage of MAD is that we can specify the costs of deviations above and below the mean differently, putting greater weight (costs) on underperformance than on outperformance. A second advantage is its computational ease. We can minimise MAD using a linear program, ie, minimise[17]

$$\frac{1}{S}\sum_{s=1}^{S}\mathrm{MAD}_s \tag{10.12}$$

with respect to MAD_s, $s = 1, \dots, S$, and $\boldsymbol{w}$ subject to the constraints

$$\begin{aligned}
\mathrm{MAD}_s - c_{\mathrm{d}}\boldsymbol{w}'(\boldsymbol{R}_s - \bar{\boldsymbol{R}}) &\geqslant 0, \quad s = 1, \dots, S\\
\mathrm{MAD}_s - c_{\mathrm{u}}\boldsymbol{w}'(\boldsymbol{R}_s - \bar{\boldsymbol{R}}) &\geqslant 0, \quad s = 1, \dots, S\\
\boldsymbol{w}'\bar{\boldsymbol{R}} &\geqslant r_{\mathrm{target}}\\
\boldsymbol{w}'\mathbf{1} &= 1\\
w_i &\geqslant 0
\end{aligned} \tag{10.13}$$

where c_{d} measures the costs of underperforming the portfolio mean, c_{u} measures the costs of outperforming the portfolio mean and r_{target} is the targeted portfolio return.

The computational mean–absolute deviation-based portfolio selection method described here offers a range of appealing properties compared with variance-based models.

1. There is no need to calculate a covariance matrix. However, this is only true if we rely on historical data for scenario generation; simulated scenarios from a parametric distribution have to be drawn using a covariance matrix.
2. Solving linear programs is much easier than mean–variance optimisation. The number of constraints ($2S + 2$ in the case of mean–absolute deviation) depends on the number of scenarios, not on the number of assets.
3. The upper bound on the number of assets in the optimal solution is related to the number of scenarios ($2S + 2$ in the case of mean–absolute deviation).

We can now use the scenarios in Table 10.5 of Appendix A (on page 298) to create an efficient frontier in mean–absolute deviation return space. Figure 10.3 plots three efficient frontiers for $c_{\mathrm{d}} = 1, 3, 5$ and $c_{\mathrm{u}} = 1$. Higher costs for underperformance will shift the frontiers to the right. This must be so by definition, but how much will the underlying portfolio weights change? Isolating the risk aspect, we will focus on the minimum-risk portfolios in Figure 10.3.

Figure 10.3 Mean–absolute deviation-efficient portfolios

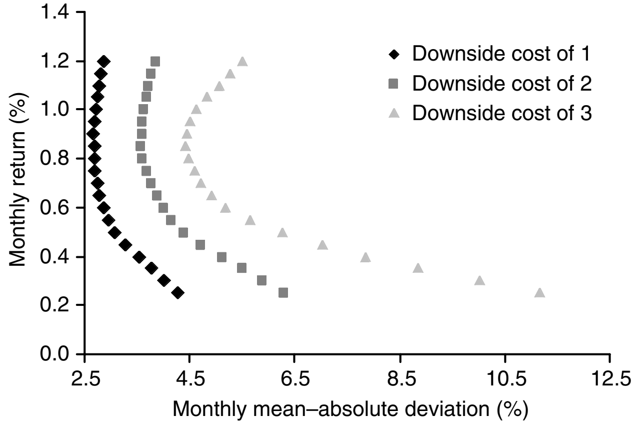


Table 10.2 Minimum-risk portfolios for different deviation costs, c_{d}

	Weights (%)			
	w_1	w_2	w_3	w_4
$c_{\mathrm{d}} = 1$	39.44	17.55	0.00	43.01
$c_{\mathrm{d}} = 3$	40.50	22.24	3.09	34.17
$c_{\mathrm{d}} = 5$	47.72	19.40	2.89	30.00

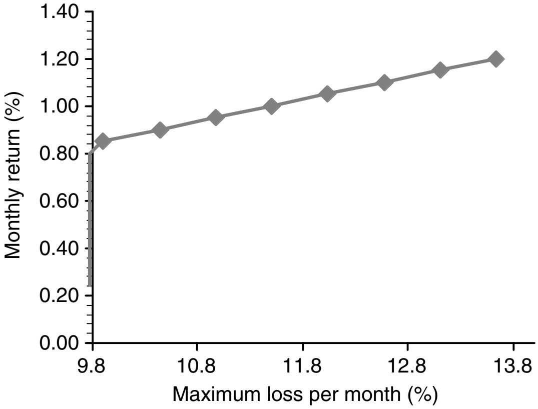
Assets 1–4 are, respectively, UK, Japanese, European and US equities.
Results calculated for return data in Table 10.5 of Appendix A.

Table 10.2 shows that the optimal position in asset 4 becomes smaller as deviation costs rise. This is not surprising as we know from Table 10.1 that US equities exhibit the most negative skewness and the highest excess kurtosis. However, as we have only 120 scenarios for our optimisation exercise, the reader will appreciate that a few observations might change the picture completely.

10.4 MINIMUM REGRET

The second objective we want to deal with is minimum regret. Suppose we obtain the scenario matrix S, either from historical returns or from a scenario simulation exercise. As in basic decision theory, we could choose a minimax criterion, ie, we might want to minimise the maximum portfolio loss, referred to as "minimising regret".[18]

Figure 10.4 Portfolio construction by minimising regret



This could be the optimal strategy for investors who have to ensure that under all means (scenarios) they never experience a particular size of loss. Focusing on extreme events has its merits if returns deviate substantially from normality or if investors are extremely risk averse.[19] Minimising the maximum loss can also be written as a linear program, ie, we maximise

$$R_{\min} \tag{10.14}$$

with respect to $\boldsymbol{w}$ and $R_{\min}$ (minimum portfolio return) subject to the following set of constraints

$$\begin{aligned} \underbrace{\boldsymbol{w}'\boldsymbol{R}_S - R_{\min}}_{\sum_{i=1}^{k} w_i(1+c+r_{is})-R_{\min}} &\geqslant 0, \quad s = 1,\dots,S \\ \boldsymbol{w}'\bar{\boldsymbol{R}} &\geqslant R_{\text{target}} \\ \boldsymbol{w}'\mathbf{1} &= 1 \\ w_i &\geqslant 0 \end{aligned} \tag{10.15}$$

where we use the same notation as in Section 10.3. The efficient frontier – in this case the geometric location of minimum regret for a given return target – is plotted in Figure 10.4.

It can be seen from the figure that the maximum loss remains constant for a wide range of return requirements. The reason for this is that the optimal solution is concentrated in 100% of asset 1 (UK equities) up to a return requirement of 0.8% per month. This in turn results from the fact that asset 1 is the asset with the smallest minimum loss in the whole sample (as the reader can see from Table 10.5 in Appendix A (on page 298)). However, this on its own would not be enough to give the above result as diversification can reduce the maximum portfolio loss if it is possible to add assets that perform well if asset 1 performs badly. Looking at the scenarios in Table 10.5 we see that all assets performed very poorly in August 1998, when all assets lost about 10% or more, leaving no room for diversification. Although a scenario-based approach allows for the fact that all assets tend to do poorly in down markets and, hence, concentrates on the least volatile asset, few practitioners would feel happy with this extreme allocation as it will maximise the regret of having chosen the wrong asset class. Portfolios with higher return requirement are constructed by shifting out of asset 1 into asset 4 (US equities), which has an average return almost as high as asset 3 (European equities) but a lower minimum return. Again, for practical purposes the extremeness of the chosen portfolios will prove to be a serious limitation.

10.5 CONDITIONAL VALUE-AT-RISK

As investors come to recognise (unfortunately, often through bitter experience, ie, unexpected losses) that risk management is an integral part of portfolio management, there is an increasing interest in value-at-risk (VaR) as a measure of investment risk. Calculations of VaR help one to assess what maximum loss – either as an absolute monetary value or in terms of return – an investment portfolio might experience with confidence β over a prespecified time period. We need to specify a level of confidence (typically 0.95, ie, 95%) as the maximum possible loss will always be 100%. Interest in VaR as a risk measure is driven by regulatory frameworks, widespread popularity and intuitive appeal. However, the last is often ill-founded, as can be seen from the following list of serious shortcomings.

1. Investors subscribing to VaR implicitly state that they are indifferent between very small losses in excess of VaR and very high losses – even total ruin. This is hardly a realistic position.
2. VaR is not sub-additive, ie, we might find that the portfolio VaR is higher than the sum of the individual asset VaRs. Suppose that we invest in a corporate bond with a 2% probability

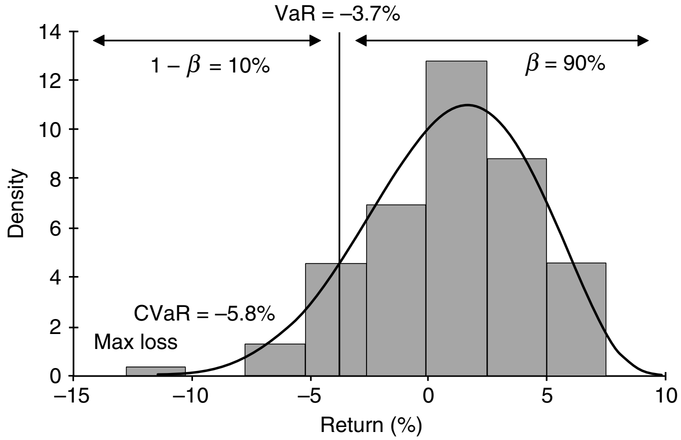
of default. With this default probability the VaR at the 95% confidence level is zero. However, if we move to a diversified bond portfolio (with zero default correlation) containing 100 bonds, the probability of at least one loss is 87% ($1 - 0.98^{100}$). Hence, portfolio optimisation using VaR would result in concentrated portfolios as the VaR measure would indicate that the concentrated portfolio was less risky.

3. VaR also has serious mathematical drawbacks. It is a non-smooth, non-convex and multi-extremum (many local minima) function that makes it difficult to use in portfolio construction. We therefore have to rely on heuristics and can never be 100% sure that the optimal solution has been found. As VaR-based scenario optimisation effectively has to count all losses higher than a moving threshold, keeping track of this changing tail, while at the same time maximising VaR, is a complicated, mixed-integer problem. Little commercially available software is available to overcome this problem.

Despite these drawbacks, investors display an astonishing interest in VaR.[20] However, there is a close relative of value-at-risk known as "conditional value-at-risk" (CVaR).[21] CVaR provides information which is complementary to that given by plain VaR in that it measures the expected excess loss over VaR if a loss larger than VaR actually occurs. Hence, it is the average of the worst $(1 - \beta)$ losses. CVaR must, by definition, be larger than VaR as it provides information about what kind of losses might hide in the tail of the distribution (hence its alternative name of "tail VaR"). Whereas VaR is a quantile, CVaR is a conditional tail expectation, ie, an expected value for those realisations that fall below the VaR threshold. Also in contrast to VaR, CVaR is sub-additive and can easily be implemented using linear programming, making it computationally very attractive. CVaR can be written as

$$R_{\text{CVaR}}(\boldsymbol{w}, \beta) = \underbrace{R_{\text{VaR}} + \underbrace{\frac{\overbrace{(1/S)\sum_{s=1}^{S} \max[R_{\text{VaR}} - \boldsymbol{w}'\boldsymbol{R}_S, 0]}^{\text{Average excess loss over all scenarios}}}{\underbrace{1 - \beta}_{\text{Probability of excess loss}}}}_{\text{Conditional excess loss}}}_{\text{Average loss if loss occurs}} \qquad (10.16)$$

Figure 10.5 Return distribution of minimum-CVaR portfolio



where R_{VaR} is the return VaR: the maximum percentage loss with confidence β. It has been shown that we can write portfolio construction by scenario optimisation using CVaR as risk measure as the linear program below,[22] where we minimise

$$R_{\text{VaR}} + \frac{1}{S}\frac{1}{1-\beta}\sum_{s=1}^{S} d_s \tag{10.17}$$

with respect to d_s, $\boldsymbol{w}$ and R_{VaR} under the constraints

$$\begin{aligned} d_s &\geqslant R_{\text{VaR}} + \boldsymbol{w}'\boldsymbol{R}_s, \quad && s = 1, \dots, S \\ d_s &\geqslant 0, && s = 1, \dots, S \\ \boldsymbol{w}'\bar{\boldsymbol{R}} &\geqslant R_{\text{target}} \\ \boldsymbol{w}'\mathbf{1} &= 0 \end{aligned} \tag{10.18}$$

Suppose we use the data in Appendix A (on page 298) as a description of a finite set of scenarios. We will examine a particular problem in which the objective is to minimise CVaR for $\beta = 0.9$.[23] The solution is given in Table 10.3, where we see that an allocation of 23.3% to asset 1, 18.5% to asset 2 and the remaining 58.2% to asset 4 yields the minimum CVaR of 9.8%. The corresponding VaR is 3.7%. The distribution of portfolio returns corresponding to this allocation and the scenario matrix $\boldsymbol{S}$ is plotted in Figure 10.5 along with the best-fitting distribution (in this case a general beta distribution) imposed on the graph. Figure 10.5 shows the maximum loss, CVaR and VaR, and it

Table 10.3 Minimum-CVaR portfolios for different confidence levels

	Confidence level				
Weights (%)	**0.75**	**0.8**	**0.85**	**0.9**	**0.95**
w_1	43.7	49.3	23.2	23.3	25.7
w_2	14.0	9.6	14.9	18.5	21.6
w_3	2.8	0.0	0.0	0.0	0.0
w_4	39.5	41.0	61.9	58.2	52.8
Risk (%)					
VaR	0.93	1.7	2.73	3.7	5.20
CVaR	3.65	4.3	4.94	5.8	6.86

Assets 1–4 are, respectively, UK, Japanese, European and US equities. Results calculated for return data in Table 10.5.

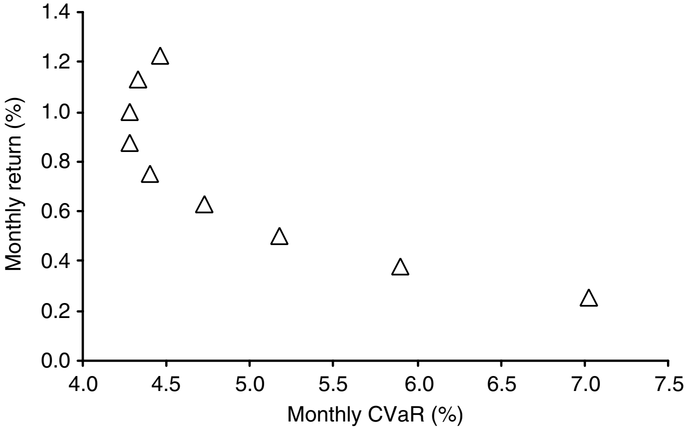
Table 10.4 Portfolios along mean CVaR frontier

	Return (%)								
Weights (%)	**0.25**	**0.38**	**0.50**	**0.63**	**0.75**	**0.88**	**1.00**	**1.13**	**1.23**
w_1	1.46	22.41	43.35	45.12	47.45	46.87	34.91	8.02	0.00
w_2	98.54	77.59	56.65	43.25	29.63	17.16	9.26	7.30	0.31
w_3	0.00	0.00	0.00	0.00	0.00	0.00	8.41	22.61	29.79
w_4	0.00	0.00	0.00	11.63	22.91	35.96	47.41	62.07	69.90
Risk (%)									
VaR	4.37	3.30	2.52	2.04	1.76	1.49	1.75	2.03	2.16
CVaR	7.03	5.89	5.17	4.73	4.40	4.28	4.27	4.33	4.46

Assets 1–4 as in previous tables.

splits the distribution into those returns smaller than VaR and those higher than VaR. It should be noted that VaR calculated from the procedure above is not the minimum VaR for a given confidence level (though in most cases it is not far from it). It is also easy to see that portfolio returns are not normally distributed, which is why we might want to use scenario optimisation. CVaR (−9.8%) is higher than VaR (−3.7%) as CVaR adds the expected excess losses to VaR. The relationship between VaR and CVaR for various confidence levels can be judged from Table 10.3. The more we move into the tail of the distribution, the closer the results both risk measures give.[24]

Figure 10.6 Mean CVaR frontier



VaR and CVaR only coincide if the tail of the distribution is cut off. Although for a normal distribution it is generally true that the two risk measures become increasingly similar as we move into the tail of a distribution (because the likelihood of large losses becomes small very fast), this is not necessarily true for non-normal distributions. There was convergence in our example as we had a limited data set that eventually truncated itself. Applying the same logic as before, we can calculate the geometric location of all mean CVaR-efficient portfolios as in Figure 10.6. Note that only a few of these portfolios are truly efficient as is the case for the other two objectives. Data for the underlying portfolios are given in Table 10.4.

As with the techniques described earlier in the chapter, linear programming tends to produce portfolios that are concentrated in a few holdings. However, the computational ease and theoretical superiority of CVaR as a risk measure make CVaR optimisation an interesting alternative to traditional portfolio construction methods.

10.6 CONCLUSION

Scenario optimisation demands considerably more effort from investors than traditional Markowitz optimisation with its widely available mean–variance optimisation software solutions. Nevertheless, scenario optimisation is worth the effort if returns are non-normally distributed or if options are involved. However, most users of scenario optimisation take a very naive approach, calculating the option payout for a given strike using the simulated asset price for that scenario. This allows only the option weight as a choice variable, whereas, for example, the value of the option strike is equally important. Moreover, this approach might generate arbitrage opportunities if the option is priced under a different distribution to that used for scenario generation. This is likely to be the case if option prices are taken from current market prices.

APPENDIX A: DATA SET FOR SCENARIO OPTIMISATION

So that the reader can replicate the calculations in this chapter, the data set on which they are based is provided in Table 10.5. The data set is a 10-year time series of monthly hedged returns (hedged into US dollars) for the four major equity markets (Europe, UK, US and Japan) plus a time series for one-month US Treasury bills, which represent the risk-free security in our analysis. The data can be viewed as a time series of returns or, alternatively, if we believe the data is drawn from a stationary distribution, we could also interpret each monthly row as a new scenario. The reader may care to check for himself that the returns are non-normally distributed, exhibiting skewness and excess kurtosis. While not a prerequisite for scenario optimisation, this nevertheless makes the technique more interesting as it can explicitly address non-normality.

APPENDIX B: SCENARIO OPTIMISATION WITH S-PLUS

We assume in this appendix that the reader is familiar with the basics of S-Plus and NuOPT for S-Plus.[25] If not, there are many excellent textbooks on this exciting modelling language that allows one to deal with complex financial modelling problems.[26] Here we will focus on mean–absolute deviation (MAD, discussed in Section 10.3) as an objective. The web page that accompanies this book offers downloads of further S-Plus functions that perform sophisticated portfolio construction and data modelling operations. Programs will include (among others) optimisations based on lower partial moments and semivariance as well as portfolio resampling. The functions will create single portfolios as well as tracing out efficient frontiers.

We start with the usual housekeeping operations – that is, we clear the workspace, include the NuOPT module, load the data, check for missing observations and store the asset names for further usage. To do this, key in the following instructions, highlight them and press F10 to execute the command

```
remove(ls(‘‘’’))
module(nuopt)
import.data(FileName=’’scenarios.xls’’,
  FileType=’’Excel’’,DataFrame=’’scenarios’’)
if(any(is.na(scenarios))==T)
  stop(‘‘no missing data is allowed’’)
asset.names<-c(dimnames(scenarios)[[2]])
```

The data frame “scenarios” contains the scenario matrix *S*. The expected format is a rectangular data matrix that contains as many columns as assets and as many rows as scenarios. The scenario matrix is assumed to be arbitrage-free. This is the only set of information we need, together with inputs for the costs of up- and downside deviations from the portfolio mean.

In the next step we write a function that defines the MAD model for NuOPT. We start by defining this function and its inputs

Table 10.5 10-year time series of monthly hedged returns for major equity markets and one-month US Treasury bills (%)

Date	UK	Japan	Europe	US	US bills
April 30 1991	0.47	−0.35	0.60	0.40	0.47
May 31, 1991	0.14	0.32	4.15	4.16	0.47
June 28, 1991	−3.72	−7.65	−3.77	−4.58	0.47
July 31, 1991	7.48	2.89	0.25	4.69	0.47
August 30, 1991	2.23	−6.16	0.97	2.43	0.45
September 30, 1991	−0.74	5.57	−2.45	−1.69	0.45
October 31, 1991	−2.21	2.97	−1.26	1.73	0.41
November 29, 1991	−6.59	−8.09	−3.56	−4.04	0.37
December 31, 1991	2.14	−1.03	0.10	11.43	0.32
January 31, 1992	2.76	−4.45	5.45	−1.66	0.32
February 28, 1992	−0.06	−5.38	2.72	1.15	0.33
March 31, 1992	−4.97	−7.97	−1.98	−1.84	0.35
April 30, 1992	9.42	−6.03	1.54	2.93	0.31
May 29, 1992	1.95	3.52	1.92	0.41	0.32
June 30, 1992	−8.24	−10.57	−6.21	−1.52	0.30

Table 10.5 (*continued*)

Date	UK	Japan	Europe	US	US bills
July 31, 1992	−5.36	−0.25	−7.25	4.17	0.27
August 31, 1992	−4.21	14.30	−5.54	−2.39	0.27
September 30, 1992	8.13	−5.11	−0.50	1.09	0.23
October 30, 1992	3.51	−1.91	1.66	0.61	0.24
November 30, 1992	4.39	3.62	3.12	3.33	0.28
December 31, 1992	3.43	−1.33	2.17	1.11	0.26
January 29, 1993	−0.97	−0.41	1.16	0.88	0.24
February 26, 1993	2.06	−1.35	5.85	1.32	0.23
March 31, 1993	1.09	11.37	1.46	2.06	0.25
April 30, 1993	−1.63	14.19	−1.42	−2.20	0.25
May 31, 1993	1.03	−0.80	1.38	2.56	0.25
June 30, 1993	2.10	−2.78	3.20	0.31	0.26
July 30, 1993	1.39	5.34	4.70	−0.10	0.25
August 31, 1993	6.05	2.29	6.41	3.62	0.26
September 30, 1993	−1.89	−3.71	−2.91	−0.98	0.25
October 29, 1993	4.35	1.95	6.68	1.76	0.26
November 30, 1993	0.23	−15.86	−2.12	−0.67	0.27
December 31, 1993	7.90	5.60	7.80	1.22	0.26
January 31, 1994	2.98	13.85	4.72	3.64	0.25
February 28, 1994	−4.27	−0.01	−4.67	−2.80	0.29
March 31, 1994	−6.73	−5.52	−2.93	−4.32	0.30
April 29, 1994	1.36	3.22	3.51	1.43	0.33
May 31, 1994	−5.02	5.38	−4.30	1.68	0.36
June 30, 1994	−1.82	−0.95	−4.88	−2.80	0.35
July 29, 1994	5.78	−1.84	5.06	3.39	0.38
August 31, 1994	5.64	0.89	1.14	3.95	0.39
September 30, 1994	−7.27	−3.47	−5.80	−2.10	0.40
October 31, 1994	2.38	0.91	0.91	2.32	0.43
November 30, 1994	−0.37	−2.77	0.74	−3.39	0.48
December 30, 1994	0.20	2.32	−0.13	1.49	0.47
January 31, 1995	−3.01	−6.31	−2.02	2.78	0.50
February 28, 1995	1.01	−6.85	0.93	3.97	0.49
March 31, 1995	4.46	−2.37	−3.01	2.70	0.49
April 28, 1995	2.59	2.47	5.11	3.16	0.49
May 31, 1995	3.57	−5.05	3.33	3.93	0.48
June 30, 1995	−0.05	−4.05	−0.33	2.48	0.46
July 31, 1995	4.96	11.98	4.04	3.42	0.47
August 31, 1995	0.76	6.59	0.62	0.03	0.45
September 29, 1995	0.78	2.00	−0.12	4.51	0.45
October 31, 1995	0.13	−1.55	−1.62	−0.03	0.46

Table 10.5 (*continued*)

Date	UK	Japan	Europe	US	US bills
November 30, 1995	3.60	5.71	2.80	4.42	0.46
December 29, 1995	1.12	7.18	2.40	1.53	0.42
January 31, 1996	2.47	2.67	5.07	3.59	0.42
February 29, 1996	−0.58	−3.01	1.07	1.06	0.42
March 29, 1996	−0.34	5.47	2.29	1.03	0.43
April 30, 1996	3.60	4.15	3.45	1.50	0.43
May 31, 1996	−2.16	−1.61	0.58	2.69	0.43
June 28, 1996	−0.92	2.39	1.60	0.61	0.43
July 31, 1996	0.45	−6.81	−4.97	−4.38	0.44
August 30, 1996	4.77	−2.29	2.33	2.26	0.44
September 30, 1996	2.26	6.38	4.87	5.62	0.42
October 31, 1996	0.51	−4.15	0.69	2.58	0.43
November 29, 1996	2.44	2.38	6.78	7.58	0.43
December 31, 1996	1.57	−4.46	2.01	−1.87	0.43
January 31, 1997	3.39	−6.04	7.88	6.87	0.43
February 28, 1997	1.25	2.15	3.94	0.73	0.44
March 31, 1997	0.36	−0.33	3.22	−4.46	0.44
April 30, 1997	2.99	6.51	2.00	6.61	0.44
May 30, 1997	3.69	2.59	2.52	5.68	0.41
June 30, 1997	−0.40	6.16	8.95	4.58	0.43
July 31, 1997	6.20	0.79	9.65	7.93	0.44
August 29, 1997	−0.78	−6.66	−9.37	−5.92	0.44
September 30, 1997	8.94	−0.71	8.28	5.24	0.43
October 31, 1997	−7.74	−9.22	−7.98	−2.71	0.43
November 28, 1997	−0.11	0.07	4.62	4.86	0.43
December 31, 1997	6.00	−3.30	5.91	1.50	0.45
January 30, 1998	5.34	6.85	5.72	1.31	0.43
February 27, 1998	6.08	0.56	7.77	7.03	0.44
March 31, 1998	3.42	−1.08	10.17	5.20	0.43
April 30, 1998	0.21	−0.85	0.12	1.19	0.42
May 29, 1998	−1.36	−0.37	4.07	−1.99	0.42
June 30, 1998	−0.74	2.01	2.53	4.31	0.42
July 31, 1998	−0.21	3.04	2.47	−0.97	0.42
August 31, 1998	−9.77	−13.21	−15.48	−13.90	0.41
September 30, 1998	−3.69	−6.01	−9.85	6.59	0.36
October 30, 1998	7.38	0.41	8.04	7.76	0.36
November 30, 1998	5.74	10.22	8.09	6.81	0.38
December 31, 1998	2.25	−4.58	3.76	5.91	0.37

Table 10.5 (*continued*)

Date	UK	Japan	Europe	US	US bills
January 29, 1999	0.08	4.15	2.55	4.32	0.37
February 26, 1999	4.64	0.17	−1.10	−2.79	0.39
March 31, 1999	2.29	14.11	2.00	4.16	0.38
April 30, 1999	4.10	5.42	4.92	3.62	0.38
May 31, 1999	−5.15	−3.22	−2.92	−2.32	0.39
June 30, 1999	1.86	9.26	3.69	5.38	0.40
July 30, 1999	−1.24	4.97	−2.84	−3.24	0.40
August 31, 1999	0.36	−4.91	3.25	−0.63	0.41
September 30, 1999	−4.00	3.79	−0.92	−2.97	0.40
October 29, 1999	3.29	2.74	5.61	6.54	0.42
November 30, 1999	4.17	2.62	7.75	2.15	0.44
December 31, 1999	4.59	6.93	12.96	6.98	0.45
January 31, 2000	−9.71	0.56	−3.57	−5.35	0.47
February 29, 2000	−0.78	0.43	10.26	−2.37	0.48
March 31, 2000	6.31	1.66	1.06	9.98	0.49
April 28, 2000	−2.26	−1.84	0.24	−3.23	0.49
May 31, 2000	1.07	−4.92	−0.99	−2.64	0.47
June 30, 2000	−0.30	5.69	−0.52	2.40	0.49
July 31, 2000	1.81	−7.47	1.02	−1.79	0.52
August 31, 2000	3.97	4.34	2.10	5.20	0.53
September 29, 2000	−4.31	−3.14	−4.68	−5.44	0.52
October 31, 2000	2.56	−4.16	2.67	−0.71	0.53
November 30, 2000	−3.43	−2.16	−5.39	−7.90	0.52
December 29, 2000	1.56	−2.55	−0.33	−0.15	0.49
January 31, 2001	0.61	1.13	1.96	3.66	0.42
February 28, 2001	−5.34	−3.24	−8.27	−8.94	0.40
March 30, 2001	−3.11	3.96	−4.84	−6.37	0.36

```
MAD.model<-function(scenarios, cost.up,
    cost.dn, r.target){}
```

The function "MAD.model" expects a scenario matrix, inputs for upside and downside deviation costs as well as a target return for the MAD optimisation. We can now start to insert the code within the { } brackets. We first calculate asset means for all assets

```
rbar<-apply(scenarios, 2, mean)
```

and count the number of scenarios

```
    n.obs<-nrow(scenarios)
```

We then set up indices to define assets and scenarios. Use “yj” to index assets and “t” to index scenarios

```
    asset<-Set()
    period<-Set()
    j<-Element(set=asset)
    t<-Element(set=period)
```

Having done that, we define parameters (which remain fixed in the optimisation, ie, they are not choice variables)

```
    R<-Parameter(scenarios, index=dprod(t,j))
    rbar<-Parameter(as.array(rbar), index=j)
    r.target<-Parameter(r.target, changeable=T)
    cost.up<-Parameter(cost.up, changeable=T)
    cost.dn<-Parameter(cost.dn, changeable=T)
```

Now we need to define the variables. As in the main text, we use weights as well as upside and downside deviations as variables

```
    x<-Variable(index=j)
    up<-Variable(index=t)
    dn<-Variable(index=t)
    up[t]>=0
    dn[t]>=0
```

The last is defined implicitly using the identity

```
    up[t]-dn[t]==Sum((R[t,j]-rbar[j])*x[j],j)
```

as “up[t]” and “dn[t]” are constrained to be non-negative. Finally, we set our objective

```
    risk<-Objective(type=''minimize'')
    risk~Sum((cost.up*up[t]+cost.dn*dn[t]),t)
                                        /(n.obs-1)
```

as well as the relevant constraints

```
    Sum(rbar[j]*x[j],j)>=r.target
    Sum(x[j],j)==1
    x[j]>=0
```

We now use the function `MAD.model` to set up a system that can be solved

```
MAD.system<-System(MAD.model, scenarios,
                  cost.up, cost.dn, r.target)
```

At this stage we can use the `show.model()` command to view the model we have just defined.

The last step is to solve the model, calculate weights as well as the value of the objective, and set up a correctly labelled weight vector

```
solution<-solve(MAD.system, trace=T)
weight<-matrix(round(solution\$variable
         \$x\$current, digit=2)*100, ncol=1)
risk<-solution\$objective
dimnames(weight)
    <-list(asset.names, ‘‘weight’’)
```

EXERCISES

1. Reproduce all graphs in this chapter with Excel using the data set provided in Table 10.5.

2. For the risk measures given in this chapter, marginal contributions to risk are often impossible to calculate in closed form. How can we best approximate these numbers? What can we do to improve the approximation?

3. Discuss the statement: "Shortfall risk measures leverage up on market momentum, hence their out-of-sample performance needs to be compared with momentum strategies."

4. Discuss the merits of full utility-based optimisation versus conditional value at risk (CVaR)? Which would you prefer? What risk aversion does CVaR imply?

[1] See Dembo (1991) as the classic reference for scenario optimisation.

[2] See Mak *et al* (1999) for a simple method for deriving confidence bands for a particular solution to a stochastic programme.

[3] See Pflug (1996) for a taxonomy of stochastic optimisation problems.

[4] We follow Grinold (1999) for the remaining part of this section. The author is indebted to R. Grinold for providing the return data set used in this chapter.

[5] See Cochrane (2001) or Bossaerts (2002) on pricing kernels.

6 See Grinold (1996).

7 Note that we do not report the sample means as we do not intend to rely on them.

8 An intuitive description of this process is found in Dembo and Freeman (1998).

9 Supposing that we have 10 years of monthly data for four time series, we can then draw random numbers from a uniform distribution ranging from one to 120. If we sample n, we take the return of the nth month across all our assets. Repeatedly sampling (with replacement, ie, August 1998 could be sampled several times within a year) and compounding returns gives us resampled multiperiod returns.

10 Bootstrapping bond returns should therefore be applied with caution as bond yields exhibit mean-reversion and duration is also a function of the interest level. (For a brief explanation of resampling, see note 9.)

11 See Lütkepohl (1991).

12 See Ziemba and Mulvey (1998).

13 See Ziemba and Mulvey (1998).

14 See Pliska (1997) for a complete definition of arbitrage opportunities.

15 The reader should note that we would also find arbitrage where $\boldsymbol{w1} = \boldsymbol{0}$, $\boldsymbol{S}^1\boldsymbol{w} \geqslant \boldsymbol{0}$ and at least one of the inequalities holds strictly. The dual problem to Equation 10.10 is min $\boldsymbol{0}'\boldsymbol{w}$ subject to $\boldsymbol{S}'\boldsymbol{\pi} = \boldsymbol{1}$, $\boldsymbol{\pi} \geqslant \boldsymbol{0}$. If the dual has a solution, then $\boldsymbol{w1}$ cannot be negative.

16 This model was introduced into the literature by Konno and Yamazaki (1991).

17 See Feinstein and Thapa (1993) or Speranza (1993) on linear optimisation in portfolio construction. It is sometimes argued that linear transaction cost constraints help to mitigate the error maximisation properties of portfolio optimisers. While it will certainly be true that the range of solutions becomes smaller, it is not clear at all how one should deal with the inherent arbitrariness in this kind of portfolio optimisation heuristic.

18 See Young (1998).

19 However, if returns are multivariate normal or investors' preferences are closely approximated by mean–variance objectives, mean–variance-based investing will show superior results.

20 See Jorion (2001) or Pearson (2002) for the application of VaR in investment management.

21 See Pflug (2000) on the relationship between VaR and CVaR.

22 See Larsen *et al* (2001).

23 The calculations have been performed using the NuOPT optimiser under S-Plus. However, they can easily be replicated using Excel. For Excel to be able to deal with the large number of constraints, the box linear problem within Solver must be ticked.

24 See also Pflug (2000) for the relationship between VaR and CVaR.

25 Extensive information on S-Plus can be found on MathSoft's home page, URL: http://www.mathsoft.com/splus/.

26 See Krause and Olson (2000) or Venables and Ripley (1997).

REFERENCES

Bossaerts, P., 2002, *The Paradox of Asset Pricing* (Princeton, NJ: Princeton University Press).

Cochrane, J., 2001, *Asset Pricing* (Princeton, NJ: Princeton University Press).

Dembo, R., 1991, "Scenario Optimization", *Annals of Operations Research* 30, pp. 63–80.

Dembo, R., and A. Freeman, 1998, *The Rules of Risk* (New York: John Wiley & Sons).

Feinstein, C., and M. Thapa, 1993, "A Reformulation of Mean Absolute Deviation Portfolio Optimization Model", *Management Science* 39, pp. 1552–3.

Grinold, R., 1996, "Domestic Grapes from Imported Wine", *Journal of Portfolio Management* 22, pp. 29–40.

Grinold, R., 1999, "Mean–Variance and Scenario-Based Approaches to Portfolio Selection", *Journal of Portfolio Management* 25, pp. 10–22.

Jorion, P., 2001, *Value at Risk*, Second Edition (New York: McGraw-Hill).

Krause, A., and Olson M., 2000, *The Basics of S and S-Plus*, Second Edition (New York: Springer).

Konno, H., and H. Yamazaki, 1991, "Mean Absolute Deviation Portfolio Optimization Model and its Application to Tokyo Stock Market", *Management Science* 37, pp. 519–31.

Larsen, N., H. Mausser and S. Uryasev, 2001, "Algorithms for Optimization of Value at Risk", Research Report no. 2001-9, University of Florida; URL: http://www.ise.ufl.edu/uryasev.

Lütkepohl, H., 1991, *Introduction to Multiple Time Series Analysis* (Berlin: Springer).

Mak, W., D. Morton and R. Wood, 1999, "Monte Carlo Bounding Techniques for Determining Solution Quality in Stochastic Programs", *Operations Research Letters* 24, pp. 47–56.

Pearson, N., 2002, *Risk Budgeting* (New York: John Wiley & Sons).

Pliska, S., 1997, *Introduction to Mathematical Finance* (Oxford: Blackwell).

Pflug, G., 1996, *Optimization of Stochastic Models* (Boston: Kluwer Academic).

Pflug, G., 2000, "Some Remarks on the Value at Risk and the Conditional Value at Risk", in S. Uryasev (ed), *Probabilistic Constrained Optimization: Methodology and Applications* (Amsterdam: Kluwer Academic), pp. 272–81.

Speranza, M., 1993, "Linear Programming Models for Portfolio Optimization", *Journal of Finance* 14, pp. 107–23.

Venables, W., and B. Ripley, 1997, *Modern Applied Statistics with S-Plus*, Third Edition (New York: Springer).

Young, M., 1998, "A Minimax Portfolio Selection Rule with Linear Programming Solution", *Management Science* 44, pp. 673–83.

Ziemba, W., and J. Mulvey, 1998, *Worldwide Asset and Liability Modelling* (Cambridge University Press).

11

Core–Satellite Investing: Budgeting Active Manager Risk

The central objective of this chapter is to show pension fund trustees how they can optimally combine the skills of both index-tracking and active fund managers. It is the most important decision after the strategic asset allocation has been derived, given the increasingly competitive nature of the pension fund market. Current practice is to use weight allocation to choose among managers (well-performing managers get higher weights while others are terminated). This chapter will show that this can be considerably more inefficient than the risk allocation method, where well-performing managers are allowed to become more aggressive, while less successful managers are moved into passive management. Instead of using the weight allocation method, the suggestion ventured is to efficiently use all available information on managers showing return histories of different lengths as truncation would lead to an increase of estimation error and hence to allocations of little practical use. Finally, the chapter offers an estimate of the loss in efficiency if the correlation structure between managers and asset classes is not properly taken into account.

Core–satellite investing is the division of funds into a passive part (the core) and an active part (one or more satellites of active managers). The main reason for this separation lies in fee arbitrage.[1] Suppose a pension fund targets 1% tracking error versus its liability-driven long-term policy benchmark. Additionally, suppose that, for the sake of simplicity, currently 100% of the pension assets are invested with a single active manager. Instead of paying active fees on 100% of its assets managed by a low (1%) tracking error manager, the pension fund could achieve the same active risk budget of 1% by investing 50% in indexed funds and 50% in a more active manager with 2% tracking error. The pension fund could then afford to pay twice as much for the more aggressive manager and still break even on fee expenditures. The idea is that while asset managers charge fees on the total volume, the biggest part of the fund is effectively indexed (dead weight).[2]

However, this often quoted calculation relies on two assumptions, which are addressed in detail later in the chapter. The first assumption is that aggressive managers can maintain the same information ratios as their less aggressive counterparts. The second is that fees are calculated as asset-based fees (current practice). If fees were calculated as lump sum payments (a fixed US dollar amount, independent of fund volume or tracking error size), mandate size would not matter and the more aggressive manager would have no fee advantage (although limits to fee arbitrage would have to eventually take effect, otherwise nothing would stop investors from moving into increasingly more aggressive core–satellite investing). It should now be clear that successful core–satellite investing has to solve a variety of problems at the heart of modern pension fund management: How many active investments should a pension fund undertake, ie, what is the proportion of satellites to core? Where (in which regions, asset classes, styles, etc) should a pension fund be active? How can multiple managers be optimally combined? The following sections demonstrate that all three questions can be solved simultaneously as a straightforward risk budgeting exercise. Risk budgeting issues are also discussed in Scherer (2000, 2001) and in Winkelmann (2000).

11.1 MATHEMATICS OF MULTIPLE MANAGER ALLOCATION: TWO-MANAGER CASE

Let us go through a simple two-manager example to study how the mechanics of optimal manager allocation work. (See Appendix A (on page 329) for a more general approach.) Active manager risk and return are given by the usual expressions

$$\begin{aligned}\alpha &= w_{\alpha_1} + w_{\alpha_2}\\ \sigma_\alpha &= (w_{\alpha_1}^2\sigma_{\alpha_1}^2 + w_{\alpha_2}^2\sigma_{\alpha_2}^2 + 2w_{\alpha_1}w_{\alpha_2}\sigma_{\alpha_1\alpha_2})^{1/2}\end{aligned}$$

where $\sigma_{\alpha i}$ (for $i = 1, 2$) denotes the tracking error of manager i and w_{α_i} is the respective weight. The marginal contribution to active risk (how much active risk – measured in tracking error – changes if we put a small additional amount of the respective manager into the portfolio) can be calculated as the first derivative of the tracking error expression

$$\frac{\mathrm{d}\sigma_\alpha}{\mathrm{d}w_{\alpha_1}} = \frac{w_{\alpha_1}\sigma^2_{\alpha_1} + w_{\alpha_2}\sigma_{\alpha_1\alpha_2}}{\sigma_\alpha} \tag{11.1}$$

Multiplying this expression by $w_{\alpha_i}/\sigma_\alpha$, we arrive at the percentage risk allocation for each manager

$$\frac{\mathrm{d}\sigma_\alpha}{\mathrm{d}w_{\alpha_1}}\frac{w_{\alpha_1}}{\sigma_\alpha} = \frac{w^2_{\alpha_1}\sigma^2_{\alpha_1} + w_{\alpha_1}w_{\alpha_2}\sigma_{\alpha_1\alpha_2}}{\sigma^2_\alpha}$$

The reader might note that this term looks like an elasticity. As already shown, the sum of these elasticities equals one as the tracking error is linearly homogeneous in terms of weight (ie, doubling weights will double tracking error). For illustrative purposes, this is repeated for the two-manager case

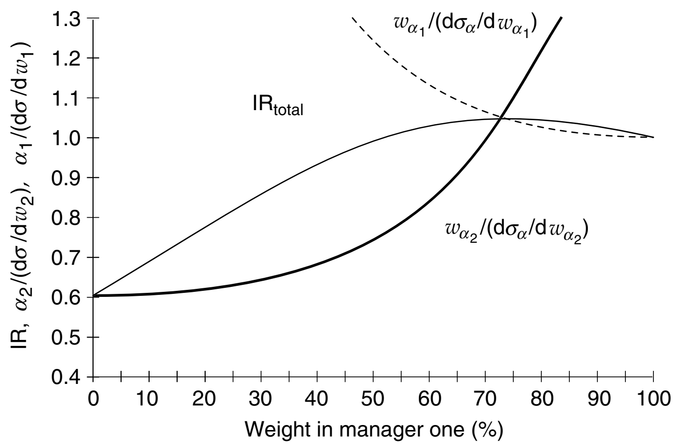
$$\begin{aligned}
&\frac{\mathrm{d}\sigma_\alpha}{\mathrm{d}w_{\alpha_1}}\frac{w_{\alpha_1}}{\sigma_\alpha} + \frac{\mathrm{d}\sigma_\alpha}{\mathrm{d}w_{\alpha_2}}\frac{w_{\alpha_2}}{\sigma_\alpha} \\
&\qquad = \frac{w^2_{\alpha_1}\sigma^2_{\alpha_1} + w_{\alpha_1}w_{\alpha_2}\sigma_{\alpha_1\alpha_2}}{\sigma^2_\alpha} + \frac{w^2_{\alpha_2}\sigma^2_{\alpha_1} + w_{\alpha_1}w_{\alpha_2}\sigma_{\alpha_1\alpha_2}}{\sigma^2_\alpha} \\
&\qquad = \frac{\sigma^2_\alpha}{\sigma^2_\alpha} = 1
\end{aligned} \tag{11.2}$$

The optimal solution for two managers can be found by trying all possible manager combinations, hence "grid searching" for the optimal solution. This involves increasing the weight of manager one (simultaneously decreasing the weight of manager two) over the whole spectrum of weights and calculating the corresponding information ratio (IR)[3] for each weight combination.[4] This allows us not only to find the optimal manager combination but also to find out more about the optimality conditions.

We assume that both managers exhibit a 5% tracking error. Manager one manages to create a 5% alpha while manager two returns an alpha of only 3%. Both managers show a correlation of 0.3. How should we optimally combine both managers? The optimality criterion is to maximise the information ratio. Figure 11.1 visualises the grid search process, plotting information ratio as well as marginal return to marginal risk ratios, as a function of manager one's weight.

The overall information ratio is maximised at a weight of about 75% for manager one and of 25% for manager two. However, Figure 11.1 also shows the ratio of marginal return to marginal risk for

Figure 11.1 Optimal manager allocation



both managers. We can see that the optimal manager allocation is characterised by an equal ratio of marginal return to marginal risk across all managers and that this ratio equals the optimal information ratio of the combined portfolio. All lines intersect at the maximum information ratio (see Appendix A (on page 329) for more details). We can therefore write

$$\frac{\alpha_1}{(w_{\alpha_1}\sigma^2_{\alpha_1} + w_{\alpha_2}\sigma_{\alpha_1\alpha_2})/\sigma_\alpha} = \frac{\alpha_2}{(w_{\alpha_2}\sigma^2_{\alpha_2} + w_{\alpha_2}\sigma_{\alpha_1\alpha_2})/\sigma_\alpha} = \mathrm{IR}^*_{\text{total}} \tag{11.3}$$

This is the core principle of multiple manager allocation. In fact, it is the core principle of any optimal portfolio decision. We can rewrite the constant relationship between performance and risk contribution by expanding both sides with $w_{\alpha_1}/w_{\alpha_2}$ and multiplying both sides by σ_α

$$\left.\begin{aligned} \frac{w_{\alpha_1}}{(w_{\alpha_1}\sigma^2_{\alpha_1} + w_{\alpha_2}\sigma_{\alpha_1\alpha_2})/\sigma^2_\alpha} &= \frac{w_{\alpha_2}}{(w_{\alpha_2}\sigma^2_{\alpha_2} + w_{\alpha_2}\sigma_{\alpha_1\alpha_2})/\sigma^2_\alpha} \\ \frac{\text{Return contribution}_1}{\text{Risk contribution}_1} &= \frac{\text{Return contribution}_2}{\text{Risk contribution}_2} \end{aligned}\right\} \tag{11.4}$$

The ratio of return to risk contribution will be the same for both managers provided they have been optimally allocated. We can also use this principle to derive the implied alphas for a given manager allocation

$$\alpha_1 = \mathrm{IR}^*_{\mathrm{total}} \frac{\mathrm{d}\sigma_\alpha}{\mathrm{d}w_{\alpha_1}} = \mathrm{IR}^*_{\mathrm{total}} (w_{\alpha_1} \sigma^2_{\alpha_1} + w_{\alpha_2} \sigma_{\alpha_1 \alpha_2}) \sigma_\alpha^{-1}$$

The implied alpha rises with tracking error (greater risk), weightings (both directly reflecting conviction about a manager's ability) and covariance (less diversifying) with other managers.

11.2 WHY MULTIPLE MANAGERS?

We have already seen in the previous section that adding a second manager increases the information ratio of the total portfolio. What would adding more managers do to the information ratio? Suppose that we can add uncorrelated managers to our portfolio and suppose for simplicity that those managers show zero correlation between them. How many managers would we add? What would happen to the information ratio? Writing out the expressions for alpha and tracking error, assuming equal weights among managers, we get

$$\mathrm{IR}_{\mathrm{total}} = \frac{\sum (1/n)\alpha_i}{(\sum (1/n)^2 \sigma^2_{\alpha_i})^{1/2}} = \frac{\bar{\alpha}}{\bar{\sigma}_\alpha} n^{1/2} \tag{11.5}$$

As the number of managers increases, the information ratio will also rise (the increase, however, is itself lessening). Naively, we might think there would be an incentive to add all available managers. What stops this happening in reality? Primarily, the average alpha in the universe is, by definition, negative.[5]

Hence, increasing the number of managers indiscriminately will still diversify active risks, but at the expense of picking up smaller or negative alphas from less capable managers, so the average will converge to zero. Moreover, it will be difficult to find more and more uncorrelated managers (availability issues). In addition, this would increase coordination, selection and monitoring costs arising from a divided manager structure.

The dominant arguments for multiple manager allocation are diversification and specialisation. Specialisation aims at the nominator in Equation 11.5. The idea is that it is unlikely that a balanced manager can provide first quartile information ratios on all decisions taken (on government bonds, credit, currency overlay, equity management, etc). Rather than accepting a diluted information ratio (diluted by less excellent decision areas), specialists should focus on selecting and should raise the average alpha in a multiple manager fund.[6] Consultants (unsurprisingly) favour this argument as it depends mainly on fund selection skills. It rests on the assumption that narrowing down the universe (making fewer prices to forecast) will result in an increased forecasting ability.[7] Specialisation in its purest form only exists if each specialist has a different benchmark reflecting a personalised special skill, appropriate to that specialist, in this particular universe.

In contrast, diversification aims at reducing risk for a given alpha level. It arises from a diversification of signals, ie, different managers create different forecasts and hence perform better, as long as the signals provide value. Forecast diversification tends to create less correlated alphas and hence assists the risk reduction in adding on another manager. In diversification's purest form, different managers are given the same benchmark. Diversification without manager selection skills, however, proves to be counterproductive; active risks would cancel out without alpha left on the table, hence the sponsor gets a passive fund at active costs.

Specialisation and diversification do not exclude each other. There is, and always will be, incentive to diversify among specialists. However, selection, monitoring and coordination costs will again limit the extent of this procedure.

So far in this chapter, it has been established that adding on individual managers can increase the active return per unit of active risk (portfolio return minus benchmark return) and hence the stability of active returns. However, would a plan sponsor choose to hold a multiple manager portfolio if in possession of all the individual managers' information? Does the separation of responsibilities still lead to an overall optimal solution? One approach to solving the problem would be "to replace decentralised management with decentralised prediction making and centralised management".[8] Effectively, this means that asset managers sell their forecasts (hence ceasing to be asset managers) and plan sponsors optimally combine these forecasts to construct a portfolio that is fully compatible with their objectives. This is a multiple adviser, rather than a multiple manager, structure.[9] However, most managers refuse to sell their predictions (not to mention that very often they do not even generate quantitative forecasts). They do so because the marginal costs of reselling their signals are very small and the revealed information is difficult to control. Moreover, they do not want to unbundle their services as a good signal is worth more to an asset manager with more under management and information can then be utilised on a greater asset base and hence earn higher fees. The problem is currently solved either by assuming it away (assuming zero correlation between active returns and asset class returns) or by imposing it away, by forcing managers to create uncorrelated active portfolios. Section 11.9 and Appendix B (on page 329) deal with this in more detail.

11.3 WHY CORE–SATELLITE?

The essence of the fee arbitrage argument supporting core–satellite investing is that partially substituting active funds for passive funds (to generate fee savings) and investing the remaining part of a portfolio into more aggressive funds (high tracking error funds at the same or moderately higher fees) will on balance result in fee savings while performance remains unaffected. High tracking error funds can be generated by the fund manager by taking a low tracking error fund and multiplying the active positions by a multiple greater than one. Alternatively, the investor could invest in an active fund and finance this investment from selling the underlying index, hence isolating the active portfolio. Repeating this several times generates any targeted leverage. However, the paradox remains that no one would buy a low tracking error fund at the same asset-based fee (fees calculated as a percentage of assets given to the manager) if a combination of passive investments and high tracking error funds yields the same active risk budget, but at lower costs. However, continuing fee arbitrage is not an equilibrium situation. If it were, every investor would move into core–satellite, reducing the amount of assets under active management. Even if fees were to rise for active funds, this would not offset the loss in revenues from a reduced amount of funds under active management (as otherwise there would not be an arbitrage situation in the first place).

If mandates were priced as lump sum payments plus performance-based fees (as costs for a given manager are likely to be unrelated to portfolio size or aggressiveness), there would no longer be fee arbitrage. Already there is a tendency towards this model as fees quoted by asset managers are very often based on breakeven dollar amounts. An asset manager will have a breakeven value (based on cost structure and equity return requirements) and scale the asset-based fee accordingly in order to generate (at least) the breakeven income. There is a caveat, however: investment managers point out that high active risk funds require different model portfolios (new model portfolios and therefore non-scalable business) and should hence be quoted at higher fees. This argument is certainly true, but it will only apply to those with active risk requirements higher than the point where the long-only constraint is binding. For all those who are not quite as aggressive, employing aggressive funds at higher costs might be an inefficient way to achieve active risk exposure.

11.4 CORE–SATELLITE: HOW MUCH SHOULD BE ACTIVE?

The answer to any “how much?” question critically relies on assumptions and opinions (how high does the investor rate active managers) as well as on preferences (risk aversion).[10]

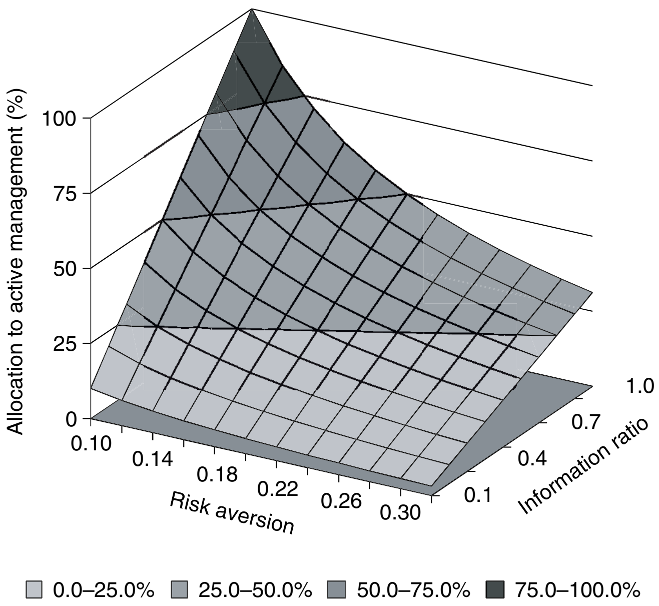
Supposing an investor can allocate funds between a passive portfolio and an active portfolio with a 5% tracking error, this would be a straightforward calculus problem; maximising the information ratio with respect to w and supposing risk aversion is 0.2 and the information ratio is 0.5, the optimal tracking error would amount to 1.25%. This solves the first-order condition of the standard utility optimisation problem[11]

$$\frac{\mathrm{d}U}{\mathrm{d}\sigma_\alpha} = \mathrm{IR} - 0.2 \cdot 2 \cdot \sigma_\alpha = 0$$

In order to generate a 1.25% tracking error, we need 25% of a 5% tracking error product. Hence, the optimal solution is 25%, as can be seen in Figure 11.2.

As we would expect, a combination of high IR and low risk aversion leads to an allocation to active management of about 100%. Observing a plan sponsor’s allocation for passive investing and knowing the risk aversion would then allow us to calculate the implied assessment of the skill (information ratio) of active managers using reverse optimisation (see Appendix A on page 329). If satellites become more aggressive (high marginal contribution to risk), core–satellite investing does not necessarily result from a fading belief in active management.

Figure 11.2 How much should be active?



11.5 WHERE TO BE ACTIVE?

The degree to which assets are actively managed is typically decided on a "top-down" basis; first choose between active and passive and then decide where active management is employed. More importantly, the optimal manager mix is found by simultaneously choosing between active and passive managers in each region (or, alternatively, style buckets) of our portfolios. This is a straightforward optimisation problem (see above), where high alpha managers get a bigger allocation (in terms of either weight or tracking error) unless they need too much risk to generate a given outperformance or unless they show high correlations with other managers. In this section we will focus on manager allocation on the basis of allocating weights.[12]

As an example, we will assume that a portfolio is split between two regions (50% each).[13] In each region we can choose between four active managers and one passive manager. We will not allow manager allocation to override regional allocations, ie, we would not concentrate our portfolio in one region purely because there are highly skilled active managers in that particular part of the world (or style bucket). Each manager initially targets a tracking error of 5%. However, their information ratio differs; managers in country one have an information ratio of one while those in country two each have an information ratio of 0.3. Active managers show correlation of active returns of 0.3 within a country and zero otherwise. The objective is to maximise alpha subject to a 1% tracking error constraint.

Given the assumptions above, the solution can be found in Figure 11.2. A considerable part of money in country one would be managed actively while there would be little active management in country two. A fraction of 36% is managed actively. The optimality condition in Equation 11.3, from the previous section, is satisfied as the ratio of marginal excess return to marginal risk contribution is equal to all assets and equals the information ratio of 1.5. This scenario can be likened to an American portfolio where the money is split between the US (a highly efficient market) and Latin America (a less efficient market). The optimal allocations support our pre-understanding that there should be more passive investing where few good managers can be found, and vice versa. As correlations are assumed to be symmetrical, with no difference in information ratio and volatility, we end up with equal weights for active managers within a given market.

Continuing this example, we can assume (holding everything else equal) that manager 1 in country 1 runs a tracking error of only 2.5%, which provides a different allocation as in Table 11.1. Most notably, manager 1 gets a much bigger weight to make up for the small tracking error. Allocation to manager 1 will be extended until the relation between marginal contribution to portfolio risk and marginal contribution to excess return equal those of the other managers. Practically, this will mean that competitive pressure will force manager 1 in country 1 to lower the asset-based fee.

However, manager allocation in country two remains unchanged as we assumed zero correlations among active returns in different regions.

11.6 RISK ALLOCATION VERSUS MANAGER ALLOCATION

So far, the examples we have used have been of the optimal manager allocation using weights rather than tracking error. This is exactly

Table 11.1 Optimal manager allocation

Manager	w (%)	IR_i	σ_i (%)	α_i (%)	$\frac{\mathrm{d}\sigma}{\mathrm{d}w_i}$ (%)	$\frac{\mathrm{d}\alpha}{\mathrm{d}w_i}\frac{w_i}{\sigma}$ (%)	$\frac{\alpha_i}{\mathrm{d}\sigma/\mathrm{d}w_i}$
Country 1, active 1	6.95	1	5	5.0	3.30	22.9	1.5
Country 1, active 2	6.95	1	5	5.0	3.30	22.9	1.5
Country 1, active 3	6.95	1	5	5.0	3.30	22.9	1.5
Country 1, active 4	6.95	1	5	5.0	3.30	22.9	1.5
Country 1, passive	22.20	0	0	0.0	0.00	0.0	—
Country 2, active 1	2.08	0.3	5	1.5	0.99	2.1	1.5
Country 2, active 2	2.08	0.3	5	1.5	0.99	2.1	1.5
Country 2, active 3	2.08	0.3	5	1.5	0.99	2.1	1.5
Country 2, active 4	2.08	0.3	5	1.5	0.99	2.1	1.5
Country 2, passive	41.67	0	0	0.0	0.00	0.0	—
Total	100.00	1.5	1	2	—	100.0	—
Country 1, active 2	6.95	1.0	5	5.0	3.30	22.9	1.5
Country 1, active 3	6.95	1.0	5	5.0	3.30	22.9	1.5
Country 1, active 4	6.95	1.0	5	5.0	3.30	22.9	1.5
Country 1, passive	15.26	0.0	0	0.0	0.00	0.0	—
Country 2, active 1	2.08	0.3	5	1.5	0.99	2.1	1.5
Country 2, active 2	2.08	0.3	5	1.5	0.99	2.1	1.5
Country 2, active 3	2.08	0.3	5	1.5	0.99	2.1	1.5
Country 2, active 4	2.08	0.3	5	1.5	0.99	2.1	1.5
Country 2, passive	41.66	0.0	0	0.0	0.00	0.0	—
Total	100.00	1.5	1	2	—	100.0	—

the problematic scenario in which a retail client might find themselves. However, institutional clients could agree with the manager on a new (more or less aggressive) tracking error target, making it possible to leave manager weights constant (weighting them equally within a region and instead choosing between managers on the basis of assigned aggressiveness).

The "ideal world" noted in Table 11.2, would have no restrictions (no short-sales limits, no country limits, etc), and both approaches (choosing either weights or tracking errors) would yield the same results, ie, the same value added. However, reality differs from the idealised world above and hence both approaches will yield different solutions. The first major difference is that there will be no room for (explicit) passive management as soon as variations in tracking error are introduced. Actually, passive management will

Table 11.2 Optimal manager allocation based on different choices

	Multiple manager allocation	
	Based on tracking error	**Based on weights**
Ideal world	Same optimal results	
Constraints on country weights only	Optimal risk allocation	Limited alpha transfer
Binding long-only constraints	Limited leverage (explicit modelling of IR slippage)	No problem

be implicit in the level of tracking error chosen as essentially a low tracking error means that only a small fraction of the fund is managed actively. Instead, all funds will be active funds with varying tracking error targets. Competent managers have a higher tracking error assigned to them while less competent managers are assigned a lower tracking error. Passive management is still present, but it is implicit (amount of dead weight in total portfolio) rather than explicit. A general side effect of the industry moving to core–satellite is that the figures of passive assets under management are probably inflated as the core–satellite approach itself just makes passive investments explicit rather than implicit.

Assuming that the short-sales constraints are not binding, it is easy to see that choosing tracking error allows greater flexibility than using weights as the constraints on physical investments are no longer binding. Suppose all good managers are located in Latin America (LA) and the LA benchmark weight happens to be 10%: choosing weights rather than tracking error restricts the alpha from LA managers as only a maximum of 10% can be invested into LA assets. However, the contribution from LA active risk to the total risk budget could be improved (increased) if the LA bets were leveraged up.

To slightly modify the previous example, assume that country 1 has a 10% benchmark weighting while country 2 has one of 90%. All other assumptions remain the same. How would a solution based on tracking error allocation (budgeting of risk) compare with a solution based on weight allocation (manager allocation)? The results can be seen in Tables 11.3 and 11.4.

Table 11.3 Optimal manager allocation (weight allocation)

Manager	w (%)	IR_i	σ_i (%)	α_i (%)	$\frac{d\sigma}{dw_i}$ (%)	$\frac{d\alpha}{dw_i}\frac{w_i}{\sigma}$ (%)	$\frac{\alpha_i}{d\sigma/dw_i}$
Country 1, active 1	2.50	1.0	5	5.0	1.19	3.0	4.2
Country 1, active 2	2.50	1.0	5	5.0	1.19	3.0	4.2
Country 1, active 3	2.50	1.0	5	5.0	1.19	3.0	4.2
Country 1, active 4	2.50	1.0	5	5.0	1.19	3.0	4.2
Country 1, passive	0.00	0.0	0	0.0	0.00	0.0	—
Country 2, active 1	6.81	0.3	5	1.5	3.23	22.0	0.5
Country 2, active 2	6.81	0.3	5	1.5	3.23	22.0	0.5
Country 2, active 3	6.81	0.3	5	1.5	3.23	22.0	0.5
Country 2, active 4	6.81	0.3	5	1.5	3.23	22.0	0.5
Country 2, passive	62.76	0.0	0	0.0	0.00	0.0	—
Total	100.00	0.9	1	1	—	100.0	—

Table 11.4 Optimal manager allocation (risk allocation)

Manager	w (%)	ϕ_i	IR_i	σ_i (%)	α_i (%)	$\frac{d\sigma}{dw_i}$ (%)	$\frac{d\alpha}{dw_i}\frac{w_i}{\sigma}$ (%)	$\frac{\alpha_i}{d\sigma/dw_i}$
Country 1, active 1	2.50	2.8	1.0	5	13.86	9.15	22.9	1.5
Country 1, active 2	2.50	2.8	1.0	5	13.86	9.15	22.9	1.5
Country 1, active 3	2.50	2.8	1.0	5	13.86	9.15	22.9	1.5
Country 1, active 4	2.50	2.8	1.0	5	13.86	9.15	22.9	1.5
Country 1, passive	0.00	—	0.0	0	0.00	0.00	0.0	—
Country 2, active 1	22.50	0.0	0.3	5	0.14	0.09	2.1	1.5
Country 2, active 2	22.50	0.0	0.3	5	0.14	0.09	2.1	1.5
Country 2, active 3	22.50	0.0	0.3	5	0.14	0.09	2.1	1.5
Country 2, active 4	22.50	0.0	0.3	5	0.14	0.09	2.1	1.5
Country 2, passive	0.00	—	0.0	0	0.00	0.00	0.0	—
Total	100.00	—	1.5	1	2	—	100.0	—

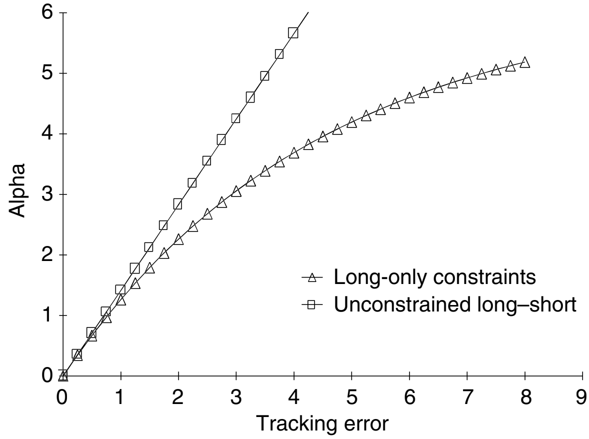
The optimisation confirms the perception we have built previously. Choosing weights (as depicted in Table 11.3) seriously constrains manager allocation because if the best managers are tied up in a small benchmark allocation, optimality conditions will not be met. While tracking error allocation correctly assigns the biggest part of the risk budget to the best managers, weight allocation fails to do so. In fact, we can see that percentage contributions to risk are equal to those in Table 11.1. This is not surprising as the benchmark does not matter in a world of perfect alpha transfer. While the information ratio for tracking error allocation is about 1.5, it only amounts to 0.9 for weight allocation. This example clearly favours tracking error-based multiple manager allocation. Leveraging investment processes indiscriminately faces some practical restrictions, as the next section contends.

11.7 LIMITATIONS OF CORE–SATELLITE INVESTING

In an ideal world, core–satellite would be 100% indexed, plus a sum of independent market-neutral long–short overlay managers (zero net market exposure with uncorrelated bets that are added on, ie, overlaid to a passive benchmark), thereby disentangling asset and manager allocation. However, this world does not comply with reality as most institutional managers face long-only constraints. We will therefore compare core–satellite investing with a so-called "enhanced indexing" approach,[14] ie, the combination of low tracking error active fund managers.[15] The ability to generate high risk-adjusted returns (high information ratio) drops as portfolios get more concentrated; this relationship is captured in Figure 11.3. Very aggressive satellite portfolios can only be generated with loss in information ratio, while there is little loss for enhanced indexing (low tracking error active management). In the chosen simulation example (a fairly typical active equity manager situation), there is virtually no loss in information ratio up to about 1% tracking error.[16] More aggressive management, however, must lead to a reduction in information ratio.

Satellites have to be very active in order to generate a target tracking error for 100% of assets while only using x% of total assets. In general, this will not work without loss in information ratio as it forces long-only managers to hold increasingly undiversified active portfolios (in order to create high tracking error). Long–short managers, however, would just scale up their position without loss in information or diversification. The extent to which this argument is true depends on the level of x chosen and the products involved (macro funds find it easier than micro funds to generate higher tracking error without loss in IR). Practical modelling will have to include an estimate for the information ratio slippage (information ratio as

Figure 11.3 Tracking error and scalability



function of tracking error). The problem then becomes maximising

$$\mathrm{IR}(\sigma_\alpha)\sigma_\alpha - f(\sigma_\alpha) - \lambda\sigma_\alpha^2 \tag{11.6}$$

with respect to risk, where $f(\sigma)$ denotes fees as a function of tracking error. A necessary condition for optimality is given by

$$\sigma_\alpha = \left(\mathrm{IR}(\sigma_\alpha) - \overbrace{\frac{\mathrm{d}f(\sigma_\alpha)}{\mathrm{d}\sigma_\alpha}}^{+}\right) \Big/ \left(2\lambda - \underbrace{\frac{\mathrm{d}\,\mathrm{IR}(\sigma_\alpha)}{\mathrm{d}\sigma_\alpha}}_{-}\right) \tag{11.7}$$

Bracketed signs indicate the derivative. An increase in tracking error will, for example, lead to a deteriorating information ratio as short-sales constraints become binding.

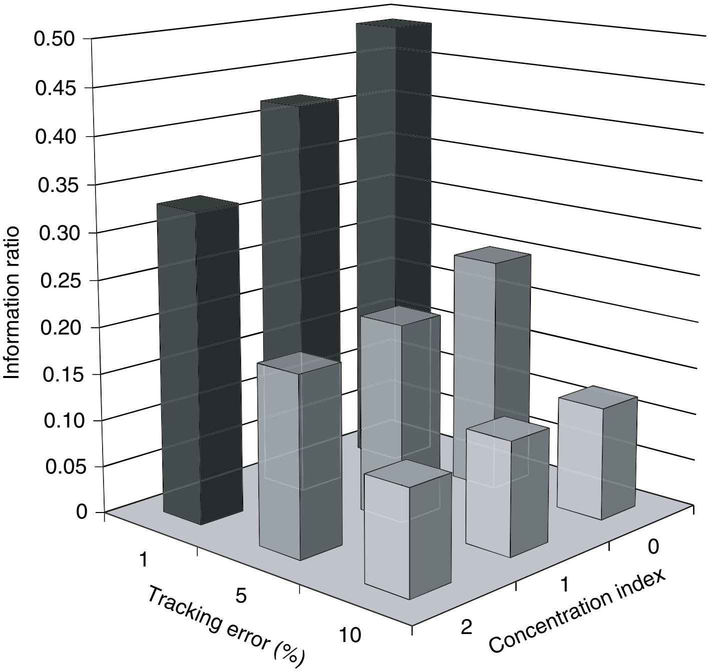
The usual optimality condition $\sigma_\alpha = \mathrm{IR}/2\lambda$ will only be obtained if fees, as well as information ratios, are independent of tracking error.

Additionally, investors are likely to create a negative size-bias as overweights (higher weightings than index weights) increasingly have to be financed from big stocks. Again, this can be seen from a Monte Carlo simulation study of manager behaviour. We choose a 100-stock benchmark with different benchmark concentration levels and also a log concentration model where zero stands for no concentration (ie, equal weighting) and concentration levels rise as the numbers get bigger.[17] Results are summarised in Figure 11.4.

Table 11.5 Core–satellite versus enhanced indexing

	Core–satellite	Enhanced indexing
Ideal world (no constraints)	Fee arbitrage favours core satellite (not a long-term equilibrium situation)	
Allocation	Equally applicable	
Binding short-sale constraint	Loss in IR for high aggressiveness levels	Manager allocation by choosing weights
Short-sale constraint plus concentrated benchmark	Negative size bias	Neutral

Figure 11.4 Benchmark concentration and information ratios



The effect of increasing benchmark concentration is most pronounced for small tracking errors. For very aggressive portfolios, there is almost no difference as the long-only constraint is already limiting enough. The biggest information ratio can be achieved for unconcentrated benchmarks and small tracking errors.

11.8 INPUT GENERATION

Portfolio optimisation is very sensitive to variations in inputs (see Chapters 4 and 6).[18] Managers with high information ratios are seen to be "attractive", and thus tend to get the highest fraction of the active risk budget. This problem is familiar to those who are involved in multiple manager allocation. Time series data of manager performance are only available in different lengths; hence estimates are prone to different estimation errors. A low information ratio might be statistically significant if there were enough data, while a high information ratio might not if it were based on a short history. Portfolio optimisers have no way of detecting this and treat both numbers with the same certainty. This section illustrates the estimation error problem and a potential solution.[19]

Suppose you have five potential value specialists in which to invest (see Table 11.6); you would be given annual returns from 1970 to 1999 (30 observations) for three funds but only half the data for two of the five funds.[20] Faced with this problem, it is common practice to truncate all time series and use the data from 1985 onwards. The results (truncated sample estimator) are shown in Table 11.7. It is obvious that this procedure throws away available information on funds one to three. Remember that the level of the information ratio determines the active–passive decision, while the dispersion determines manager allocation. But, can this be improved upon?

Alternatively, we can use maximum likelihood techniques that incorporate the correlation between fund series. For the long series (funds one to three ranging from 1970 to 1999), αs are still found by averaging over the 30-year period. Risks are also calculated using the long data series as these are the maximum likelihood estimates. For the short series, we first compare the αs for the longer time series before and after 1985. The difference is scaled according to the sensitivities (see Chapter 6 on Bayesian methods for a detailed explanation of methodology) found by two multiple regressions of the returns for funds four and five on the remaining funds one to three in Table 11.6. As returns have been different in both periods – higher up to 1985 – information ratios are adjusted upwards.[21]

However, the dispersion of information ratios is still relatively high, and well above the true information ratio of 0.5 each, as estimation error is not accounted for. Using an uninformative prior to reflect

Table 11.6 Manager data with different time lengths

Year	Fund 1	Fund 2	Fund 3	Fund 4	Fund 5
1970	0.63	0.33	0.59	n.a.	n.a.
1971	−1.83	−0.34	−1.14	n.a.	n.a.
1972	1.75	2.68	3.31	n.a.	n.a.
1973	2.08	2.81	0.33	n.a.	n.a.
1974	−0.79	1.21	0.81	n.a.	n.a.
1975	3.88	4.21	4.72	n.a.	n.a.
1976	3.88	4.83	4.82	n.a.	n.a.
1977	1.42	0.00	1.42	n.a.	n.a.
1978	2.15	2.00	2.06	n.a.	n.a.
1979	1.85	1.59	2.42	n.a.	n.a.
1980	1.13	−0.72	0.96	n.a.	n.a.
1981	2.95	2.97	2.52	n.a.	n.a.
1982	0.32	−0.71	−0.41	n.a.	n.a.
1983	5.87	6.69	5.43	n.a.	n.a.
1984	1.23	0.32	−0.73	n.a.	n.a.
1985	0.73	1.32	0.71	1.63	2.91
1986	2.63	2.47	2.45	2.88	4.00
1987	0.62	−0.51	0.44	−0.32	−0.21
1988	0.31	−2.26	0.74	−0.80	−1.35
1989	−1.16	−0.90	−1.24	−0.86	−0.64
1990	1.09	−0.24	1.10	−0.07	0.52
1991	−2.17	−0.90	−0.45	−1.63	−1.64
1992	1.93	2.25	2.93	4.23	3.57
1993	3.75	5.13	2.93	4.23	5.67
1994	−0.88	0.10	−0.06	−0.97	−0.56
1995	2.22	1.10	1.79	2.07	1.26
1996	3.01	2.96	1.63	1.72	2.36
1997	−2.69	−3.26	−3.39	−3.71	−2.00
1998	−2.38	−1.83	−0.65	−1.23	−1.87
1999	1.64	1.36	1.25	2.82	0.88

estimation error, we get the estimates labelled “Bayes”. Taking estimation error into account will not change the means, as means take on the role of expected values and there is no uncertainty about an expected value. However, risks are increased considerably by scaling down information ratios. It is worth noting that risk increases are higher for the short series (higher estimation error). Bayesian adjustment can help in a multiple manager allocation exercise in order to reflect different time series lengths and estimation error, and hence arrive at more realistic active–passive allocation decisions.

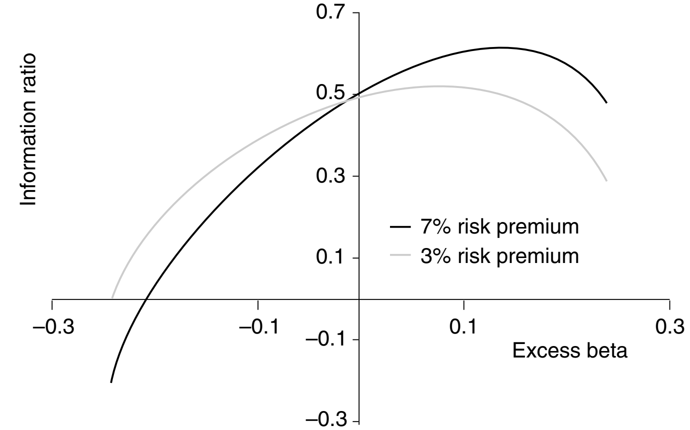
Table 11.7 Estimation results

	Fund 1	Fund 2	Fund 3	Fund 4	Fund 5
			Truncated sample estimator		
α	0.58	0.45	0.68	0.67	0.86
σ	2.03	2.28	1.94	2.32	2.38
IR_i	0.28	0.20	0.35	0.29	0.36
			Maximum likelihood		
α	1.17	1.16	1.24	1.42	1.58
σ	1.93	2.19	1.71	2.49	2.38
IR_i	0.61	0.53	0.73	0.57	0.66
			Bayes		
α	1.17	1.16	1.24	1.42	1.58
σ	2.24	2.55	1.98	2.90	2.78
IR_i	0.52	0.45	0.63	0.49	0.57

11.9 ALLOCATING BETWEEN ALPHA AND BETA BETS

So far, the manager allocation problem has been seen in isolation from the asset allocation problem; we neither took into account the interaction between active returns and asset class returns nor looked at the incentives of an asset manager to engage in structural bets. A structural bet is a position that is designed to collect the long-term (average) risk premium of an asset class. It is designed to stay on the portfolio, irrespective of market conditions and effectively changes a client's benchmark. Suppose an asset manager has got the objective to maximise risk-adjusted active returns, with a tracking error target of 5% (assuming there are no further restrictions), and a market-neutral portfolio construction information ratio of 0.5. The risk premium for taking on equity risk is 7% and equity volatility is 20%. How should the manager then allocate the 5% risk budget to maximise active returns? Introducing the possibility of structural bets, active returns can now be generated by two competing sources. One is to permanently increase equity exposure (structural beta bet); the other is to take market-neutral (beta-neutral) stock bets, so-called "alpha bets". For every unit of active risk created by the structural equity bet, one unit of alpha risk is crowded out for a given risk

Figure 11.5 Structural risk and information ratio



budget. What would be the optimal allocation of alpha and beta bets for a single manager? Figure 11.5 plots the total information ratio relative to the excess beta (portfolio beta minus one) for the above example.[22]

The information ratio increases in excess beta. However, there is a limit where the active portfolio becomes very concentrated in beta bets and the information ratio falls again. If the excess beta is zero, then we arrive at the information ratio for pure stock picking. In an environment where structural bets are attractive (high risk premium) there is a considerable incentive for the portfolio manager to take on structural risk and crowd out alpha risk. For 7% risk premium, an excess beta of 0.15 would be optimal, as in the example above. Figure 11.5 also shows that the incentive to crowd out alpha bets would be lower in a low risk premium environment. Collecting the risk premium by effectively changing the benchmark would be optimal from the myopic view of a single manager, but what would be the consequence for the correlation between managers and asset classes, and between managers themselves? If two managers take the same structural bet, their correlation would rise. In fact, if all active risk is concentrated in the structural bet, it would reach one.[23] The same would happen to the correlation between active returns of any given manager and their benchmark. Increasing correlations have a negative impact on total risk. How would the plan sponsor's total risk exposure change if a manager kept the tracking error limit the same but substituted alpha bets for structural bets? Figure 11.6 plots total risk against excess beta. We can see that taking structural bets effectively changes the sponsor's benchmark (hence violating risk tolerance). Instead of a risk close to 20%, the sponsor might end with 25% total risk even though tracking error remains unchanged.

We have looked at the problem mostly from a single manager view so far, but what would be the consequences in a multiple manager framework? Suppose a plan sponsor has to determine optimal asset allocation (β bets) and multiple manager allocation (α bets). Alpha bets yield active returns, while beta bets yield a risk premium over cash. The combined covariance matrix of asset class risk premiums (here indexed by $\beta : \boldsymbol{\Omega}_{\beta\beta}$) and active manager returns (indexed by $\alpha : \boldsymbol{\Omega}_{\alpha\alpha}$) is given by

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{\beta\beta} & \boldsymbol{\Omega}_{\beta\alpha} \\ \boldsymbol{\Omega}_{\alpha\beta} & \boldsymbol{\Omega}_{\alpha\alpha} \end{bmatrix}$$

We also assume that αs are perfectly transportable and that the sponsor does not face any add-up constraints (neither managers nor asset classes have to add up to one, short positions in both are allowed). The total problem can then be written down as

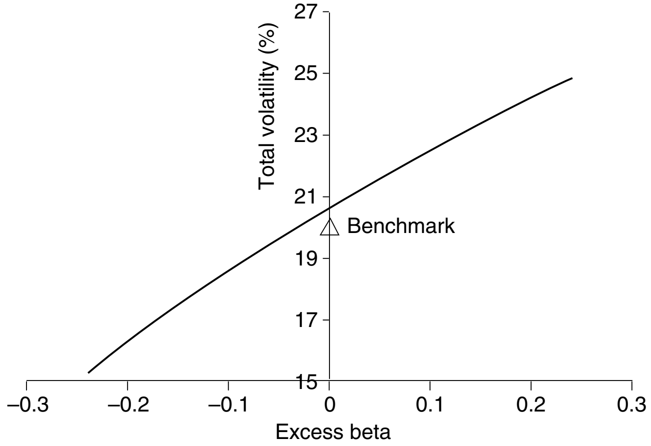
$$\begin{aligned} U &= \boldsymbol{w}'\boldsymbol{\mu} - \frac{1}{2\lambda}\boldsymbol{w}'\boldsymbol{\Omega}\boldsymbol{w} \\ \boldsymbol{w}' &= \begin{bmatrix} \boldsymbol{w}_{\beta} & \boldsymbol{w}_{\alpha} \end{bmatrix} \\ \boldsymbol{\mu}' &= \begin{bmatrix} \boldsymbol{\mu}_{\beta} & \boldsymbol{\mu}_{\alpha} \end{bmatrix} \\ \boldsymbol{\Omega} &= \begin{bmatrix} \boldsymbol{\Omega}_{\beta\beta} & \boldsymbol{\Omega}_{\beta\alpha} \\ \boldsymbol{\Omega}_{\alpha\beta} & \boldsymbol{\Omega}_{\alpha\alpha} \end{bmatrix} \end{aligned}$$

Expanding this expression we get

$$U = \underbrace{\boldsymbol{w}'_{\alpha}\boldsymbol{\mu}_{\alpha} - \frac{1}{2\lambda}\boldsymbol{w}'_{\alpha}\boldsymbol{\Omega}_{\alpha\alpha}\boldsymbol{w}_{\alpha}}_{\text{Multiple manager problem}}$$
$$+ \underbrace{\boldsymbol{w}'_{\beta}\boldsymbol{\mu}_{\beta} - \frac{1}{2\lambda}\boldsymbol{w}'_{\beta}\boldsymbol{\Omega}_{\beta\beta}\boldsymbol{w}_{\beta}}_{\text{Asset allocation problem}} - \underbrace{\frac{1}{\lambda}\boldsymbol{w}'_{\beta}\boldsymbol{\Omega}_{\alpha\beta}\boldsymbol{w}_{\alpha}}_{\text{Interaction term}}$$

Current practice has been to (often implicitly) assume the off-diagonal matrixes' correlation to be zero and to solve the problem separately.[24] If this were true, both problems could be separated, and

Figure 11.6 Impact of structural bets on total volatility



we would ignore the interaction term without losing efficiency. Solutions for separate optimisation are given below (see the appendix for more details)

$$\boldsymbol{w}^{*}_{\beta,\mathrm{sep}} = \lambda \boldsymbol{\Omega}^{-1}_{\beta\beta}\boldsymbol{\mu}_{\beta} \quad \text{and} \quad \boldsymbol{w}^{*}_{\alpha,\mathrm{sep}} = \lambda \boldsymbol{\Omega}^{-1}_{\alpha\alpha}\boldsymbol{\mu}_{\alpha}$$

However, it becomes apparent that the same risk aversion drives beta and alpha bets. Applying different risk aversions in the separated problems is inconsistent with the total problem and yields a sub-optimal solution.[25]

Alternatively, an asset manager could run a partial optimisation where optimal asset allocation weights are independently derived in a first step and manager weights are calculated in a second step, taking into account cross-correlations between managers and asset classes. The partial optimisation can be shown to lead to

$$\boldsymbol{w}^{*}_{\alpha,\mathrm{par}} = \lambda(\boldsymbol{\Omega}^{-1}_{\alpha\alpha} - \boldsymbol{\Omega}^{-1}_{\alpha\alpha}\boldsymbol{\Omega}_{\alpha\beta}\boldsymbol{\Omega}^{-1}_{\beta\beta}\boldsymbol{\mu}_{\beta})$$

Substituting the solution vectors back into the utility function of the total optimisation problem, we get a positive expression for the loss in utility

$$U(\boldsymbol{w}^{*}_{\mathrm{par}}) - U(\boldsymbol{w}^{*}_{\mathrm{sep}}) = \tfrac{1}{2}\lambda\boldsymbol{\mu}^{\mathrm{T}}_{\beta}\boldsymbol{\Omega}^{-1}_{\beta\beta}\boldsymbol{\Omega}_{\beta\alpha}\boldsymbol{\Omega}^{-1}_{\alpha\alpha}\boldsymbol{\Omega}_{\alpha\beta}\boldsymbol{\Omega}^{-1}_{\beta\beta}\boldsymbol{\mu}_{\beta}$$

This difference would approach zero if correlations became small. Separate overlays of active managers, ignoring cross-correlations between managers and asset classes, yield to a loss in utility. This is identical to other asset allocation problems (currency management, balanced portfolio management or asset–liability management) where it is also vital not to ignore correlation. However, while correlations between bonds and currencies are given and cannot be changed by the investor, this is not the case in manager allocation exercises. Plan sponsors who impose an orthogonality constraint on active returns do not have to worry about these results.[26]

11.10 CONCLUSIONS

The key factor driving the common move into core–satellite investing is fee arbitrage. As long as fee savings outweigh reduced information ratios, then the core–satellite approach is preferable to enhanced indexing. By definition, arbitrage is not an equilibrium concept. Explicit allocations to passive investments depend on whether investors allocate between managers on the basis of weights or tracking error. Allocations based on tracking error facilitate alpha transfer and hence are in many cases superior to allocations based on choosing manager weights.

Successful core–satellite investing, as a risk budgeting exercise, will have to deal with implementation issues like the incorporation of estimation error and the optimal treatment of time series covering different data periods and lengths. (In line with the arguments raised in Chapter 3, we used an optimisation framework to derive the risk budgets.) Even though core–satellite investing is the route currently favoured by many investors, enhanced indexing should not be overlooked as an alternative. Multiple manager allocations very often follow a two-step process; the first step derives the optimal strategic asset allocation, while the second step allocates across managers given the first-step asset allocations. This procedure will lead to a loss in portfolio efficiency unless sponsors have imposed orthogonality conditions on manager alphas.

APPENDIX A: MULTIPLE MANAGER MATHEMATICS

Any fund can be decomposed into a combination of an index fund (core) and an active portfolio (long–short satellite).[27] As we can see, every benchmarked investor is already investing in a particular core–satellite product, namely one core (benchmark) and one satellite (zero investment long–short portfolio). Scaling up active bets using a scaling factor (ϕ) will not increase the information content of any given portfolio; the difference will be leverage, not information.[28] Measures of skill, like the information ratio, are independent of leverage and will hence show no difference between an aggressive implementation of a particular investment strategy and a less aggressive implementation. Two active portfolios, which only differ in leverage, will have a correlation of one. Let us write down the return (alpha) for a core–satellite product with a total allocation (w%) in active funds (w_i) as

$$\left.\begin{aligned} \alpha = \sum_i w_i \phi_i \alpha_i = \sum_i w_i \alpha_i, \qquad \sum_i w_i = w \\ w_i \phi_i = w_i, \qquad \phi_i \geqslant 0 \end{aligned}\right\} \tag{11.8}$$

where α_i denotes the alpha (portfolio minus benchmark return) of the ith manager. A more aggressive implementation of the same strategy will yield a higher expected alpha. Investors have two ways of allocating between multiple managers; they can allocate either on the basis of assigning weights to active managers (weight allocation) or on the basis of aggressiveness, ie, tracking error (risk allocation). Both ways do not necessarily lead to the same results, as we will show later. Risks in a core–satellite approach can be calculated as

$$\sigma_\alpha = \left(\sum_i \sum_j \omega_{\alpha_i} \omega_{\alpha_j} \sigma_{\alpha_i} \sigma_{\alpha_j} \right)^{1/2} \tag{11.9}$$

where $\sigma_{\alpha_i \alpha_j}$ denotes the covariance between managers' alphas. As tracking error is linearly homogeneous in either weights or scaling factors, we can use a simple Euler equation approach to decomposing tracking error. The risk budget (percentage of active risk attributed to the ith manager) can then be calculated as a sum of elasticities adding up to 1

$$\varepsilon_i = \frac{\delta \sigma}{\delta \omega_i} \frac{\omega_i}{\sigma}, \qquad \sum_i \varepsilon_i = 1$$

Optimising the risk budget and, hence, simultaneously determining the active allocation (how much should be active?) and the allocation between managers (where to be active?) are now a straightforward optimisation exercise.[29] Minimise total portfolio risk subject to an alpha constraint[30]

$$\left.\begin{aligned} \text{min: } \bar{\sigma} &= \left(\sum_i \sum_j \omega_i \omega_j \sigma_{ij}\right)^{1/2} \\ \text{subject to: } \alpha &= \sum_i \omega_i \alpha_i \end{aligned}\right\} \tag{11.10}$$

with regard to either w, weight allocation, or ϕ, risk allocation. In order to facilitate the mathematics, we change to a matrix representation of Equation 11.10 and write the corresponding Lagrangian together with the first-order conditions

$$L = \boldsymbol{\omega}' \boldsymbol{\Omega}_{\alpha\alpha} \boldsymbol{\omega} - \gamma(\boldsymbol{\omega}' \boldsymbol{\alpha} - \alpha_{\text{target}}) \tag{11.11 a}$$

$$\frac{\mathrm{d}L}{\mathrm{d}\boldsymbol{\omega}} = \boldsymbol{\Omega}_{\alpha\alpha} \boldsymbol{\omega} - \gamma \boldsymbol{\alpha} = 0 \tag{11.11 b}$$

$$\frac{\mathrm{d}L}{\mathrm{d}\gamma} = \boldsymbol{\omega}' \boldsymbol{\alpha} - \alpha_{\text{target}} = 0 \tag{11.11 c}$$

where $\boldsymbol{\omega}$ and $\boldsymbol{\alpha}$ are column vectors of manager weights and manager alpha, and $\boldsymbol{\Omega}_{\alpha\alpha}$ denotes the covariance matrix of alphas. From Equation 11.11 b we get $\boldsymbol{\omega} = \boldsymbol{\Omega}_{\alpha\alpha}^{-1} \boldsymbol{\alpha} \gamma$. Inserting this into Equation 11.11 c yields

$$\gamma = \frac{1}{\boldsymbol{\alpha}' \boldsymbol{\Omega}_{\alpha\alpha}^{-1} \boldsymbol{\alpha}} \alpha_{\text{target}}$$

The optimal weight vector can now be found as

$$\boldsymbol{\omega}^* = \frac{\boldsymbol{\Omega}_{\alpha\alpha}^{-1} \boldsymbol{\alpha}}{\boldsymbol{\alpha}' \boldsymbol{\Omega}_{\alpha\alpha}^{-1} \boldsymbol{\alpha}} \alpha_{\text{target}} \tag{11.12}$$

Substituting this into the formula for tracking error variance gives us the expression for the minimal risk level for the given alpha target: $\sigma_\alpha^2 = (\boldsymbol{\alpha}' \boldsymbol{\Omega}_{\alpha\alpha}^{-1} \boldsymbol{\alpha})^{-1} \alpha_{\text{target}}^2$. Comparing this with the expression for the Lagrange multiplier, we write $\gamma = \sigma_\alpha^2 / \alpha_{\text{target}}$. Hence, Equation 11.11 b becomes

$$\boldsymbol{\Omega}_{\alpha\alpha} \boldsymbol{\omega} = \frac{\sigma_\alpha^2}{\alpha_{\text{target}}} \boldsymbol{\alpha}$$

$$\frac{\boldsymbol{\Omega}_{\alpha\alpha} \boldsymbol{\omega}}{\sigma_\alpha} \frac{\alpha_{\text{target}}}{\sigma_\alpha} = \boldsymbol{\alpha}$$

$$\frac{\boldsymbol{\Omega}_{\alpha\alpha} \boldsymbol{\omega}}{\sigma_\alpha} \left(\frac{\alpha_{\text{target}}}{\sigma_\alpha}\right) = \boldsymbol{\alpha}$$

The individual optimality conditions for multiple manager allocation can be read as elements of the above vectors[31]

$$\frac{\mathrm{d}\sigma_\alpha}{\mathrm{d}\omega_{\alpha_i}}\left(\frac{\alpha_{\text{target}}}{\sigma_\alpha}\right) = \alpha_i \tag{11.13}$$

as

$$\frac{\mathrm{d}\sigma_\alpha}{\mathrm{d}\boldsymbol{\omega}} = \begin{bmatrix} \mathrm{d}\sigma_\alpha/\mathrm{d}\omega_{\alpha_1} \\ \vdots \\ \mathrm{d}\sigma_\alpha/\mathrm{d}\omega_{\alpha_n} \end{bmatrix} = \frac{\mathrm{d}(\boldsymbol{\omega}'\boldsymbol{\Omega}_{\alpha\alpha}\boldsymbol{\omega})^{1/2}}{\mathrm{d}\boldsymbol{\omega}}$$

$$= \tfrac{1}{2}(\boldsymbol{\omega}'\boldsymbol{\Omega}_{\alpha\alpha}\boldsymbol{\omega})^{-1/2}2\boldsymbol{\Omega}_{\alpha\alpha}\boldsymbol{\omega} = \frac{\boldsymbol{\Omega}_{\alpha\alpha}\boldsymbol{\omega}}{\sigma_\alpha}$$

It also follows from Equation 11.13 that

$$\frac{\alpha_i}{\alpha_j} = \frac{\mathrm{d}\sigma}{\mathrm{d}\omega_{\alpha_i}} \bigg/ \frac{\mathrm{d}\sigma}{\mathrm{d}\omega_{\alpha_j}} \tag{11.14}$$

Optimal risk budgeting implies that the ratio of marginal contribution to alpha and marginal contribution to risk equals the information ratio of the total multiple manager portfolio for every single fund (assuming no short-sale constraints).

APPENDIX B: MULTIPLE MANAGER ALLOCATION AND CORRELATION STRUCTURE

Using the same technique as in the currency hedging appendix to Chapter 3, the partial optimisation solution can be found by solving

$$U = \boldsymbol{w}'_\alpha\boldsymbol{\mu}_\alpha - \frac{1}{2\lambda}\boldsymbol{w}'_\alpha\boldsymbol{\Omega}_{\alpha\alpha}\boldsymbol{w}_\alpha + \boldsymbol{w}'_\beta\boldsymbol{\mu}_\beta - \frac{1}{2\lambda}\bar{\boldsymbol{w}}'_\beta\boldsymbol{\Omega}_{\beta\beta}\bar{\boldsymbol{w}}_\beta + \frac{1}{\lambda}\bar{\boldsymbol{w}}'_\beta\boldsymbol{\Omega}_{\alpha\alpha}\boldsymbol{w}_\alpha \tag{11.15}$$

with respect to $\boldsymbol{w}_\alpha$ leaving $\bar{\boldsymbol{w}}_\beta = \lambda\boldsymbol{\Omega}_{\beta\beta}^{-1}\boldsymbol{\mu}_\beta$ fixed. The result is given

$$\boldsymbol{w}^*_{\alpha,\text{par}} = \lambda(\boldsymbol{\Omega}_{\alpha\alpha}^{-1} - \boldsymbol{\Omega}_{\alpha\alpha}^{-1}\boldsymbol{\Omega}_{\alpha\beta}\boldsymbol{\Omega}_{\beta\beta}^{-1}\boldsymbol{\mu}_\beta)$$

Simultaneous optimisation, taking the interaction term into account, yields

$$\begin{aligned}\boldsymbol{w}^*_{\beta,\text{sim}} &= \lambda((\boldsymbol{\Omega}_{\beta\beta} - \boldsymbol{\theta}\boldsymbol{\Omega}_{\alpha\alpha}\boldsymbol{\theta}')^{-1}\boldsymbol{\mu}_\beta \\ &\qquad - (\boldsymbol{\Omega}_{\beta\beta} - \boldsymbol{\theta}\boldsymbol{\Omega}_{\alpha\alpha}\boldsymbol{\theta}')^{-1}\boldsymbol{\Omega}_{\alpha\alpha}^{-1}\boldsymbol{\Omega}_{\alpha\beta}\boldsymbol{\mu}_\alpha) \qquad (11.16)\\ \boldsymbol{w}^*_{\alpha,\text{sim}} &= \lambda\boldsymbol{\Omega}_{\alpha\alpha}^{-1}\boldsymbol{\mu}_\alpha - \boldsymbol{\Omega}_{\alpha\alpha}^{-1}\boldsymbol{\Omega}_{\alpha\beta}\boldsymbol{w}^*_\beta \qquad (11.17)\end{aligned}$$

Alpha and beta bets are mutually corrected for their communality. Those managers who take structural bets (systematic exposure to marketwide influences also shared by the benchmark) will end up with a lower allocation, as if these would remain undetected. However, in practice, this is rarely done. By moving from the separate to the partial solution, we can now check whether there is an increase in utility. We only have to substitute the solution vectors into the utility function to achieve

$$U(\boldsymbol{w}^{p}) - U = (\boldsymbol{w}^{S}) = \tfrac{1}{2}\lambda\boldsymbol{\mu}_{\beta}^{\mathrm{T}}\boldsymbol{\Omega}_{\beta\beta}^{-1} - \boldsymbol{\Omega}_{\beta\alpha}\boldsymbol{\Omega}_{\alpha\alpha}^{-1}\boldsymbol{\Omega}_{\alpha\beta}\boldsymbol{\Omega}_{\beta\beta}^{-1}\boldsymbol{\mu}_{\beta} \qquad (11.18)$$

As long as $\boldsymbol{\Omega}_{\beta\beta}^{-1} - \boldsymbol{\Omega}_{\beta\alpha}\boldsymbol{\Omega}_{\alpha\alpha}^{-1}\boldsymbol{\Omega}_{\alpha\beta}\boldsymbol{\Omega}_{\beta\beta}^{-1}$ is positive definite, the utility difference (quadratic form) will be positive. We proved that separate optimisation of manager and asset allocation leads (in an otherwise unconstrained context) to a utility loss, as long as manager alpha and asset allocation show nonzero correlation. In the case of zero correlation, all conditional distributions become unconditional and separation of alpha and beta bets will still result in the optimal solution. However, positive correlation between alpha and beta bets might even arise when individual alphas are orthogonal (uncorrelated) with their respective benchmarks. Suppose a manager with a pure government benchmark takes a “diversification bet” into emerging market debt. While the alpha from this structural bet would be roughly orthogonal to the government bond benchmark (low correlation of emerging market debt), there would be a unitary correlation between this bet and the emerging market part of the strategic allocation (beta bet).

EXERCISES

1. Replicate Figure 11.5. You can use the programming language of your choice or try poptools.xla in Excel.

2. Suppose you run a fund of (hedge) funds. You can invest into three uncorrelated strategies with information ratios of 3, 2 and 1. Would equal weighting be advisable? What is the optimal weighting given the above information?

[1] The availability of very aggressive products in the form of hedge funds might also serve as a catalyst. Sometimes multiple manager allocation is referred to as an additional driving force. However, this can also be achieved in a traditional framework.

[2] See Freeman (1997).

[3] The information ratio is defined as the ration of active return, ie, portfolio return minus benchmark return. Practitioners misleadingly use active return and alpha synonymously. While this is wrong – the term alpha is reserved for risk-adjusted active returns – we will

use this convention until otherwise noted. The reader should keep in mind, though, that the information ratio as a measure of risk-adjusted out-performance, as defined above, can easily be gained by investing into a riskier asset. The resulting active return would arise from collecting a risk premium rather than from superior information.

4 Grid searching is only feasible for problems with up to three managers (changing weights for two managers and treating the third manager as residual). Higher-order problems can be solved using commercially available optimisers.

5 This follows directly from Sharpe's arithmetic of active management (Sharpe 1991).

6 However, specialisation does not come without costs. Breaking up a balanced mandate into regional bonds and equity mandates would make timing and asset allocation decisions much more difficult to implement, unless an overlay manager is employed, where this is legally feasible.

7 Examples for specialisation are asset class-specific specialists like currency overlay manager, dedicated fixed income specialist, value manager, etc.

8 Sharpe (1991, p. 219).

9 See Rosenberg (1977).

10 See Sorensen *et al* (1998) for a similar approach.

11

$$\max U = \mathrm{IR} \cdot \sigma_\alpha - \lambda \sigma_\alpha^2, \qquad \frac{\mathrm{d}U}{\mathrm{d}\sigma_\alpha} = \mathrm{IR} - 2\lambda\sigma_\alpha$$

12 See Di Bartolomeo (1999) on issues in multi-manager investing.

13 Regional division might not always be the best division. A plan sponsor will try to save fees in supposedly very efficient markets and will be prepared to employ active managers in less developed asset classes. Hence, they might make large stocks the core and invest into active smaller company management. (This could still be done on a capitalisation equivalent basis, so that no discrepancies between the sum actual benchmarks and strategic benchmarks occur.)

14 Enhanced indexing is defined as low tracking error active management. Sometimes the term is used for quantitatively driven products with tight risk controls. A very good treatment of enhanced indexing can be found in Di Bartolomeo (2000).

15 This section draws heavily on the results of Grinold and Kahn (1999, Chapters 14 and 15) and puts them into the context of core–satellite investing.

16 The simulation methodology is directly taken from Grinold and Kahn (1999, p. 419ff). Suppose there is an investment universe of 200 stocks with a residual volatility of 20% each. The stocks are equally weighted in the benchmark. The portfolio manager adjusts their return expectation for each stock only once a year. Their information coefficient (correlation between forecast and realisation) is 0.1. We perform a simulation and obtain a series of alphas for each stock: $\alpha_i = 0.1 \cdot 20\% \cdot \varepsilon_i$, with $\varepsilon_i \sim N(0, 1)$. For each simulation, we construct an optimal portfolio subject to the short-selling restriction.

17 See Grinold and Kahn (1999, pp. 428–30) for a description of the methodology.

18 See Chopra and Ziemba (1993), Michaud (1998), Britten-Jones (1999) and Chapter 4 of this book. Although estimation error is already dealt with in Chapters 4 and 6, it is worthwhile repeating this issue in a multiple manager context.

19 The proposed methodology stems from Stambaugh (1997).

20 The data is generated by drawing random active returns with 2% active risk and 1% expected active return. Correlation is assumed to be 0.8. Hence, the true information ratio is 0.5. All return realisations before 1985 for funds four and five are deleted to arrive at the test data set.

21 The underlying assumption is that returns are drawn from a stationary distribution.

22 The graph can be easily constructed by writing the information ratio as a function of excess beta only

$$\mathrm{IR}(\beta) = \frac{(\beta - 1)(R - r) + \mathrm{IR}(\bar{\sigma}^2 - (\beta - 1)^2 \sigma_{\mathrm{m}}^2)^{1/2}}{\bar{\sigma}_\alpha}$$

where $R - r$ denotes the risk premium on equities, σ_{m} denotes equity volatility and $\bar{\sigma}_{\alpha}$ the prefixed tracking error.

23 The correlation between two managers can be calculated as

$$\rho_{ij}\frac{(\beta_i - 1)(\beta_j - 1)\sigma_{\mathrm{m}}^2}{\bar{\sigma}_{\alpha i}\bar{\sigma}_{\alpha j}}$$

24 Muralidhar (2001) also stresses the importance of taking into account manager versus manager and manager versus asset class correlation.

25 It is worth noting that this condition is not met in practice. Suppose an investor holds a benchmark with a volatility of 15 and an excess return over cash of 3. This results in a marginal rate of substitution (trades off return against variance) of 0.00667. Being 99% invested yields a variance of 220.52 (2.97 risk premium), while a market exposure of 101% yields a variance of 229.52 (3.03 risk premium). Dividing the difference in return by the difference in variance gives the marginal rate of substitution between return and variance. Investing into an active manager with 5% tracking error and 1% alpha (which many would already deem undesirable) reveals a marginal rate of substitution of 0.04, which is many times bigger than the one implied in the asset allocation choice.

26 Orthogonality means that active returns and benchmark returns are uncorrelated. In a CAPM world this can also be obtained by imposing beta-neutrality.

27 Strictly speaking, this is not quite true as the theoretical loss of a zero investment long–short portfolio can become infinite (a short position in a rising stock can theoretically lead to unlimited losses).

28 For the use of scaling factors in asset management, see also Lee (2000, Chapter 5).

29 See Chow and Kritzman (2001), who uses VaR to describe risk budgets. In a normally distributed world, both approaches are equivalent.

30 Baierl and Chen (2000) optimise with more difficult constraints.

31 See Grinold and Kahn (1999, p. 134).

REFERENCES

Baierl, G., and P. Chen, 2000, "Choosing Managers and Funds: How to Maximise Your Alpha without Sacrificing Your Target", *Journal of Portfolio Management* 26, Autumn, pp. 47–53.

Britten-Jones, M., 1999, "The Sampling Error in Estimates of Mean–Variance Efficient Portfolio Weights", *Journal of Finance* 54, pp. 655–71.

Chopra, V., and W. Ziemba, 1993, "The Effects of Errors in Means, Variances and Covariances on Optimal Portfolio Choice", *Journal of Portfolio Management* 19, Winter, pp. 6–11.

Chow, G., and M. Kritzman, 2001, "Risk Budgets", *Journal of Portfolio Management* 27, Winter, pp. 56–60.

Di Bartolomeo, D., 1999, "A Radical Proposal for the Operation of Multi-Manager Investment Funds", Northfield Information Services, Inc.

Di Bartolomeo, D., 2000, "The Enhanced Index Fund as an Alternative to Enhanced Index Equity Management", Northfield Information Services, Inc.

Freeman, J., 1997, "Investment Deadweight and the Advantages of Long Short Investment Management", *VBA Journal*, pp. 11–14.

Grinold, R., and R. Kahn, 1999, *Active Portfolio Management*, Second Edition (New York: McGraw-Hill).

Lee, W., 2000, *Theory and Methodology of Tactical Asset Allocation*, Frank J. Fabozzi Series (New York: John Wiley & Sons).

Michaud, R. O., 1998, *Efficient Asset Management: A Practical Guide to Stock Portfolio Optimization and Asset Allocation* (New York: Oxford University Press).

Muralidhar, A., 2001, "Optimal Risk Adjusted Portfolios with Multiple Managers", *Journal of Portfolio Management* 27, Spring, pp. 97–104.

Rosenberg, B., 1977, "Institutional Investment with Multiple Portfolio Managers", *Proceedings of the Seminar on the Analysis of Security Prices*, University of Chicago, pp. 55–160.

Scherer, B., 2000, "Preparing the Best Risk Budget", *Risk* 13, pp. 30–2.

Scherer, B., 2001, "Tracking Error and Funding Assumptions", *Journal of Asset Management* 2, pp. 235–40.

Sharpe, W. F., 1991, "The Arithmetic of Active Management", *Financial Analysts Journal* 47, pp. 7–9.

Sorensen, E., K. Miller and V. Samak, 1998, "Allocating between Active and Passive Management", *Financial Analysts Journal* 54, pp. 18–31.

Stambaugh, R., 1997, "Analysing Investments Whose Differences Vary in Length", *Journal of Financial Economics* 45, pp. 285–331.

Winkelmann, K., 2000, "Managing Active Risk at the Total Fund Level", in L. Rahl (ed), *Risk Budgeting*, pp. 39–64 (London: Risk Books).

12

Benchmark-Relative Optimisation

12.1 TRACKING ERROR: SELECTED ISSUES

Tracking error measures the dispersion (volatility) of active returns (portfolio return minus benchmark return) around the mean active return. It is designed as a measure of relative investment risk and was introduced into the academic arena in the early 1980s.[1] Since then it has become the single most important risk measure in communications between asset manager, client and consultant.

Benchmark-relative investment management has often been rationalised from either a risk perspective (a benchmark anchors the portfolio in risk–return space and thus gives sponsors confidence in what their money is invested in and what risk this investment carries), or a return perspective (claiming that it is easier to forecast relative returns than total returns). However, the return argument looks spurious; to say that forecasting relative distances is possible whereas forecasting absolute distances is not ignores the fact that the two are closely related: a total distance is the same as the benchmark distance times a relative distance. A more plausible argument that is made is that the estimation error is smaller for relative than for absolute return forecasts.

12.1.1 Tracking error and time aggregation

Assume that decisions on the weight of asset i relative to the benchmark b are made on the basis of relative risk premiums, $r_i - r_b$. It can then be shown that the volatility of relative returns will be smaller than the volatility of absolute returns of asset i if[2]

$$\operatorname{var}(r_i - r_b) < \operatorname{var}(r_i) \Leftrightarrow \sigma_i^2 + \sigma_b^2 - 2\rho\sigma_i\sigma_b < \sigma_i^2 \Leftrightarrow \rho > \frac{1}{2}\left(\frac{\sigma_b}{\sigma_i}\right)$$

Tracking error (TE) is calculated in its simplest standard deviation-based *ex post* form as

$$\sigma_\mathrm{a} = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(r_{\mathrm{a},t} - \bar{r}_\mathrm{a})^2} \tag{12.1}$$

Small deviations from the mean active return, $\bar{r}_\mathrm{a}$, carry little weight, while large deviations are increasingly punished as a result of the square function.[3] This is also called a "return-based" estimate. The *ex ante* TE is calculated using the familiar quadratic form

$$\sigma_\mathrm{a} = \sqrt{\boldsymbol{w}'_\mathrm{a}\boldsymbol{\Omega}\boldsymbol{w}_\mathrm{a}} \tag{12.2}$$

where $\boldsymbol{w}_\mathrm{a}$ denotes a $k \times 1$ vector of active positions. It is therefore also called a "weight-based" estimate.

Now suppose that we construct monthly portfolios with the same *ex ante* expected TE as in Equation 12.2 in each month. How can we aggregate this into an annual number? Assuming for simplicity that annual returns are additive ($r_{\mathrm{a},12} = \sum_{t=1}^{12} r_{\mathrm{a},t}$), we can calculate the volatility of annual active returns from[4]

$$\mathrm{var}\left(\sum_{t=1}^{12} r_{\mathrm{a},t}\right) = \mathrm{var}(r_{\mathrm{a},1} + \cdots + \mathrm{var}(r_{\mathrm{a},12}) = 12\sigma_\mathrm{a}^2 \tag{12.3}$$

Suppose we have calculated the *ex ante* TE for each month to be $1\%/\sqrt{12}$ and assume also that, even though we made no forecast error for the monthly TE, the realised TE is still different from 1%. What could have caused this discrepancy? How realistic is the above procedure? Essentially, Equation 12.3 assumes period by period active returns to be serially independent. However, as most tactical positions tend to stay on an account for a prolonged time and as factors driving active returns tend to be trending (serially correlated), this will also cause active return to be autocorrelated.[5] This means that every negative (positive) active return is now more likely to be followed by another negative (positive) active return, which creates more dispersion in cumulative active returns than if they were not autocorrelated. At this point we can borrow from the logic of variance ratio tests, which test for serial correlation in asset returns.[6] If the "true" annual TE is higher than the annualised TE and the variance ratio

$$\frac{(\text{True tracking error})^2}{(\text{Annualised tracking error})^2} = \frac{\mathrm{var}(\sum_{i=1}^{12} r_i)}{12\sigma_\mathrm{a}^2} = \text{Variance ratio}$$

is significantly greater than one, we could conclude that returns show positive serial correlation. Now we can turn the idea around and calculate a leverage factor (variance ratio) that has to be applied to single-period volatilities if returns over multiple periods are autocorrelated

$$\underbrace{\operatorname{var}\left(\sum_{1}^{12} r_{\mathrm{a},t}\right)}_{\substack{\text{Actual annual}\\ \text{volatility}}} = \underbrace{\left(1 + 2\sum_{i=1}^{11}\left(1-\frac{i}{12}\right)\rho_i\right)}_{\text{Variance ratio}} \underbrace{12\sigma_{\mathrm{a}}^2}_{\substack{\text{Naive}\\ \text{annual}\\ \text{volatility}}} \tag{12.4}$$

where ρ_i is the ith-order autocorrelation coefficient. If autocorrelations were zero, we could simply multiply the monthly volatility by 12 to arrive at annual volatility. We can use Equation 12.4 first to calculate the "fudge factor" and then to adjust the naive volatility estimate. Though it might look tempting to use this fudge factor, it should also be kept in mind that autocorrelations are not easier to forecast than TE – in fact they are notorious for their variability.

It is important to note that TE has no link with performance.[7] It would be incorrect to conclude that a fund with a TE of 0.5% per annum would (assuming normality) have a one in six chance of underperformance by more than 0.5%. The obvious example is a fund that underperforms/outperforms its benchmark by exactly 10bp every month: as the volatility of a constant is zero, we would find that the fund and benchmark drifted apart considerably over the long run, yet the TE using the conventional calculation given above would remain zero.[8] This is equivalent to saying that we cannot have an educated discussion about a normal distribution by considering only the dispersion parameter.[9] The core problem is that TE is used to calculate performance ranges under the silent assumption that average active performance is zero, while at the same time realised TE is measured and estimated using the non-zero sample mean. It would be more realistic to reconcile both assumptions by also basing TE estimates on the zero-mean assumption. The average of squared active returns would yield a higher TE and yet be more consistent with the assumption of zero active return.

12.1.2 Misestimation of tracking error

Ex ante TE (usually weight-based estimates calculated from multi-factor risk models) is an estimate of the volatility of future active returns around the mean expected active return. It is heavily used in determining the relative riskiness of active strategies by managers who are given client-specific risk limits. However, if *ex ante* TE is compared with its *ex post* counterpart (the volatility of historic active returns), the measures are usually found to differ. There are many reasons why Equation 12.2 might provide a poor forecast for Equation 12.1, and it is not possible (or advisable) to counter every one of them. It should be emphasised that although TE underestimation has been the main focus of recent interest, overestimation is in many cases almost equally likely. In this section we review the main causes of TE misestimation, the four most important of which are as follows.

1. *Sampling error* (too few data).[10] As the true TE is unknown, we have to rely on realised historical data, but since historical data is by definition limited, we will (under the assumption that model parameters do not change) always misestimate the true TE. However, the usual maximum likelihood methods will yield unbiased estimates, so we cannot claim that sampling error leads to over- or underestimation of TE. This will change for optimised portfolios, where estimation error always leads to underestimation of expected TE (see Chapter 4).
2. *Modelling error* (non-stationary data, or wrong process assumptions).[11] TE estimates based on the stationarity assumption will fail if volatilities or correlations are time-varying.[12] Although conditional risk (multifactor) models are certainly more suited to dealing with time-varying moments, over- or underestimation is equally likely. Rising correlations lead – other things being equal – to lower TEs, whereas falling correlations result in increased ones. As sudden changes in correlation are often the result of exogenous shocks (for example, the increased correlation between all bond markets after the surprise Federal Reserve interest rate hike in early 1994 and the reduced correlation between high- and low-grade fixed-income instruments after the Russian crisis of 1998), they are by definition difficult to forecast. Modelling error also arises if we assume uncorrelated period by period returns (in individual time series or risk factors) when in fact they are not uncorrelated. The most prominent example here is the autocorrelation in growth style factor returns during the TMT (technology, media and telecom) bubble. Attempts to fix this problem by

estimating the correlation structure should be treated with caution as this calls for forecastability of returns, making risk management a secondary issue. The additional noise from ill-suited forecasts will increase misestimation rather than reducing it.

3. *In-sample optimisation.*[13] Portfolio optimisation using historical data is essentially an in-sample optimisation exercise. As portfolio optimisers "over-adjust" to the data – exploiting every minor, statistically insignificant and economically implausible opportunity – they will be disappointed by out-of-sample performance.[14]

4. *Constant weight assumption.*[15] TE estimates are based on the assumption that portfolio weights stay fixed, while in fact they do not. Adding uncertainty about weights will unambiguously tend to an underestimation of TE.[16] However, the importance of this effect is considered differently.[17]

12.1.3 Tracking error optimisation

TE optimisation can be solved within the framework already set out in Chapter 1. We set $\boldsymbol{A} = \mathbf{1}'$ and $\boldsymbol{b} = \mathbf{0}$ in Equation 1.3, where $\mathbf{1}$ is a $k \times 1$ vector of 1s. The utility maximisation now becomes

$$\text{Utility} \approx \boldsymbol{w}'\boldsymbol{\mu} - \frac{1}{2\lambda}\boldsymbol{w}'\boldsymbol{\Omega}\boldsymbol{w} \quad \text{subject to } \mathbf{1}'\boldsymbol{w} = \mathbf{0} \tag{12.5}$$

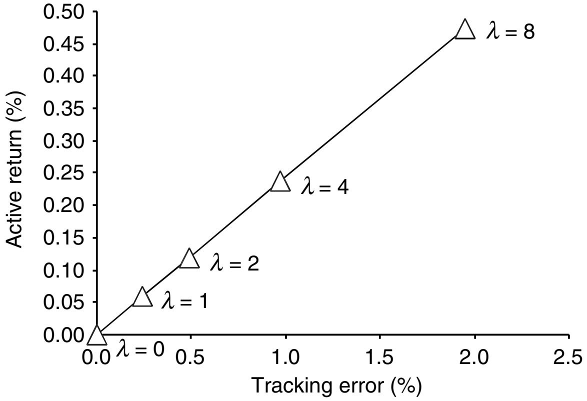
with the familiar looking solution

$$\boldsymbol{w}_{\mathrm{a}}^{*} = \lambda\boldsymbol{\Omega}^{-1}(\boldsymbol{\mu} - \mathbf{1}(\mathbf{1}'\boldsymbol{\Omega}^{-1}\mathbf{1})^{-1}\mathbf{1}'\boldsymbol{\Omega}^{-1}\boldsymbol{\mu}) \tag{12.6}$$

Note that it is the self-financing constraint (over- and underweights must sum to zero) that gives the vector of portfolio weights a new interpretation. A change in aggressiveness (λ) will scale the optimal weights up or down, but it will leave the slope of the TE-efficient frontier unchanged. Again, we will use a simple numerical example to illustrate this. Suppose we are given expected returns, covariance matrix and benchmark weights on a given asset universe as

$$\boldsymbol{\Omega} = \begin{bmatrix} 0.04 & 0.0252 & 0.0036 & 0.0024 \\ 0.0025 & 0.0441 & 0.0038 & 0.0025 \\ 0.0036 & 0.0038 & 0.0036 & 0.0014 \\ 0.0024 & 0.0025 & 0.0014 & 0.0016 \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} 4 \\ 3 \\ 2 \\ 1 \end{bmatrix} \tag{12.7}$$

Figure 12.1 Active opportunities and aggressiveness



For a risk tolerance of one we optimally invest into

$$\boldsymbol{w}_{\mathrm{a}}^{*} = \boldsymbol{\Omega}^{-1}(\boldsymbol{\mu} - \mathbf{1}(\mathbf{1}'\boldsymbol{\Omega}^{-1}\mathbf{1})^{-1}\mathbf{1}'\boldsymbol{\Omega}^{-1}\boldsymbol{\mu}) = \begin{bmatrix} 0.65\% \\ -0.008\% \\ 3.92\% \\ -4.58\% \end{bmatrix} \tag{12.8}$$

The active frontier (geometric location of active opportunities) is plotted in Figure 12.1; it starts at the origin, where a zero tolerance for active risk implies zero active weights and, hence, a 100% investment in the passive benchmark.

All portfolios along the active frontier are perfectly correlated as the active portfolio weights differ only by a linear scaling, which depends on the investor's aggressiveness. This also means that all portfolios contain exactly the same information. Note that so far, no information on benchmark weights has been needed. As long as we can go short (and face no additional constraints), the optimal active portfolio is independent of a particular benchmark. This results in perfect scalability (an investment process is deemed scalable if a given set of active positions – ie, the model portfolio – can be implemented on various client benchmarks with little or no effort), which is a necessary requirement for modern investment processes as it allows new business to be managed at low marginal cost, the marginal costs of transforming an already produced model portfolio to a different benchmark being almost zero.

12.2 TRACKING ERROR EFFICIENCY VERSUS MEAN–VARIANCE EFFICIENCY

Most practitioners feel uncomfortable with benchmark portfolios and TE constraints as they feel that the TE straightjacket is detrimental to overall portfolio efficiency. Despite long being scorned, academia finally provided help.[18] Suppose that a manager is given a benchmark and they forecast that over the next year some assets will deliver the same return as the benchmark but at significantly reduced total risk (this is the same as saying that the benchmark is not efficient). While this is an attractive opportunity in total risk–return space, it is much less so in TE–active return space. The investment is expected to yield no return in excess of the benchmark return but would expose the portfolio manager to active risk. A fixed-income investor with a domestic government bond benchmark has no incentive to invest into currency-hedged international fixed-income instruments (which give the same expected return under the assumption of negligible transaction costs and a zero risk premium on currency risk) as they would take on TE without compensation in the form of increased returns. This leads to a paradox. Plan sponsors hire a manager precisely because they think the manager has better information, but the manager cannot exploit this information advantage under the imposed TE constraint.

We will now look more formally at the key question considered in this section: how do TE-efficient portfolios compare with Markowitz-efficient portfolios in mean–variance space, ie, are TE-efficient portfolios also mean–variance-efficient? Optimal active portfolios were constructed in Section 12.1. In this section these active portfolios will be added to a given benchmark portfolio and the resulting long portfolios in mean–variance space will be compared.

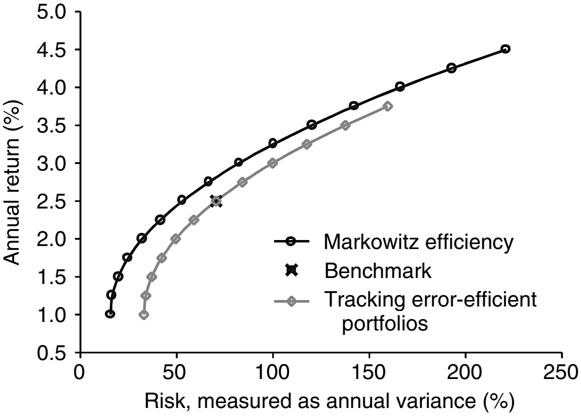
Suppose that the benchmark portfolio is given as

$$w_{\mathrm{b}} = (0.25 \quad 0.1 \quad 0.55 \quad 0.1)'$$

We can now add the active (zero-investment portfolios) from Figure 12.1 to this benchmark portfolio and compare mean–variance-efficient portfolios with TE-efficient portfolios in mean–variance space. In general (with the exception of a mean–variance-efficient benchmark, in which case the equality sign would hold) we can write

$$w_{\mathrm{TE}} = w_{\mathrm{b}} + \lambda w_{\mathrm{a}}^{*} \neq w_{\mathrm{mv}}^{*} \tag{12.9}$$

Figure 12.2 Comparison of TE-efficient and mean–variance-efficient portfolios



Portfolios constructed according to Equation 12.9 are plotted in Figure 12.2. The frontier of all TE-efficient portfolios passes through the (inefficient) benchmark and at all levels of return is dominated by the mean–variance-efficient frontier. Even though, in our setting, the portfolio manager has all the information required to reach the Markowitz efficient frontier, the imposed TE constraint makes doing so unattractive to him. Every TE-efficient portfolio will lie exactly the same distance away from a mean–variance-efficient portfolio with the same expected return as the benchmark portfolio itself.[19]

It would appear, then, that there is an independent portfolio that, if added to a TE-efficient portfolio, would result in a Markowitz efficient portfolio. We see from Figure 12.2 that this portfolio would have to have the following properties (among others):[20]

Zero-investment portfolio. This is necessary to guarantee 100% investment (weights in a zero-investment portfolio sum to zero and the portfolio's variance equals the distance between the two frontiers).

Zero return. Adding the independent portfolio to a TE-efficient portfolio must leave the return expectation constant.

Zero covariance with Markowitz efficient portfolios. If this were not so the distance would not remain constant as the covariance terms would destroy additivity.

Assuming TE optimisation under an add-up constraint, weights for this add-on portfolio, $\boldsymbol{w}_z$, can in general be derived as

$$\boldsymbol{w}_z = \frac{1}{d}[(be - af)\boldsymbol{\Omega}^{-1}\mathbf{1} + (bf - ce)\boldsymbol{\Omega}^{-1}\boldsymbol{\mu} + d\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}] \tag{12.10}$$

where $a = \boldsymbol{\mu}'\boldsymbol{\Omega}^{-1}\boldsymbol{\mu}, b = \boldsymbol{\mu}'\boldsymbol{\Omega}^{-1}\mathbf{1}, c = \mathbf{1}'\boldsymbol{\Omega}^{-1}\mathbf{1}, d = ac - b^2, e = \boldsymbol{\mu}'\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}$, $f = \mathbf{1}'\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}$ is the $k \times 1$ vector of asset covariances with the benchmark assets (for all k assets). Using the data in our example, we obtain

$$\boldsymbol{w}_z = \frac{1}{d}\left[(be - af)\boldsymbol{\Omega}^{-1}\mathbf{1} + (bf - ce)\boldsymbol{\Omega}^{-1}\boldsymbol{\mu} + d\boldsymbol{\Omega}^{-1}\begin{bmatrix}147\\130\\34\\18\end{bmatrix}\right]$$

$$= \begin{bmatrix}0.09\\0.12\\-0.52\\0.31\end{bmatrix} \tag{12.11}$$

We can now use $\boldsymbol{w}_z$ to check whether the data fulfil the three requirements outlined above. As required, the return of $\boldsymbol{w}_z$ is equal to

$$\begin{bmatrix}0.09\\0.12\\-0.52\\0.31\end{bmatrix}\begin{bmatrix}4\\3\\2\\1\end{bmatrix} = 0.09 \cdot 4 + \cdots + 0.31 \cdot 1 = 0$$

with covariance of zero with respect to all benchmark portfolios

$$\begin{bmatrix}0.09\\0.12\\-0.52\\0.31\end{bmatrix}\begin{bmatrix}0.04 & 0.0252 & 0.0036 & 0.0024\\0.0025 & 0.0441 & 0.0038 & 0.0025\\0.0036 & 0.0038 & 0.0036 & 0.0014\\0.0024 & 0.0025 & 0.0014 & 0.0016\end{bmatrix}\begin{bmatrix}-0.015\\-0.018\\0.082\\0.095\end{bmatrix} = 0$$

and variance (equal to the distance between the frontiers) of

$$\begin{bmatrix}0.09\\0.12\\-0.52\\0.31\end{bmatrix}\begin{bmatrix}0.04 & 0.0252 & 0.0036 & 0.0024\\0.0025 & 0.0441 & 0.0038 & 0.0025\\0.0036 & 0.0038 & 0.0036 & 0.0014\\0.0024 & 0.0025 & 0.0014 & 0.0016\end{bmatrix}\begin{bmatrix}0.09\\0.12\\-0.52\\0.31\end{bmatrix} = 17.35$$

In general we can express the relationship between Markowitz efficient portfolios and mean–variance-efficient portfolios with equal return as

$$w'_{\mathrm{mv}}\Omega w_{\mathrm{mv}} - w'_{\mathrm{TE}}\Omega w'_{\mathrm{TE}} = w'_{z}\Omega w_{z} \qquad (12.12)$$

In practice, however, some institutional portfolios also include so-called "beta" constraints into TE optimisation, ie, plan sponsors constrain the sensitivity of portfolio returns with respect to benchmark returns. They do this in an attempt to avoid leverage (or deleverage) as part of portfolio construction, setting the sensitivity to 1. This is also called "beta neutrality". Effectively, this will destroy the scalability of an investment process as the optimal active positions now depend on the benchmark, whereas previously they did not.[21] This analysis shows that, if TE optimisation is imposed on the active manager, setting the appropriate benchmark is important because there is no room to improve on an inefficient benchmark later.

12.3 BENCHMARK-RELATIVE PORTFOLIO CONSTRUCTION: PRACTICAL ISSUES

Practitioners who undertake benchmark-relative portfolio construction will almost always face two questions.

1. What exactly should they be forecasting – should they attempt to forecast risk premiums or deviations from equilibrium returns?
2. Does portfolio optimisation really add anything? Is it not enough just to have good forecasts?

We address the first question with an example. Suppose that equities offer a fair risk premium (equilibrium risk premium) of 5% per annum over cash with a volatility of 20% and cash yields of 5%. Assume we have varying skill with an information coefficient, IC, ranging from 0.0 to 0.2. The information coefficient captures the correlation between forecast return and return realisation.[22] An asset manger can forecast total returns, R_i, with an expected mean of 10%; risk premiums, $R_i - c$, with an expected mean of 5%; or deviations from the fair risk premium, $R_i - c - \pi_i$, with an expected mean of 0%. Each time the forecast variable is positive, equities are overweighted by 50% relative to cash. What are the expected active returns (portfolio return minus benchmark return) from these strategies? Monte Carlo simulation yields the results given in Table 12.1, where active returns are measured in basis points (bp).

Table 12.1 Skill, forecast and performance

	Skill, measured as information coefficient		
Forecast	**IC = 0**	**IC = 0.1**	**IC = 0.2**
R_i	100	160	240
$R_i - c$	50	120	200
$R_i - c - \pi_i$	0	80	160

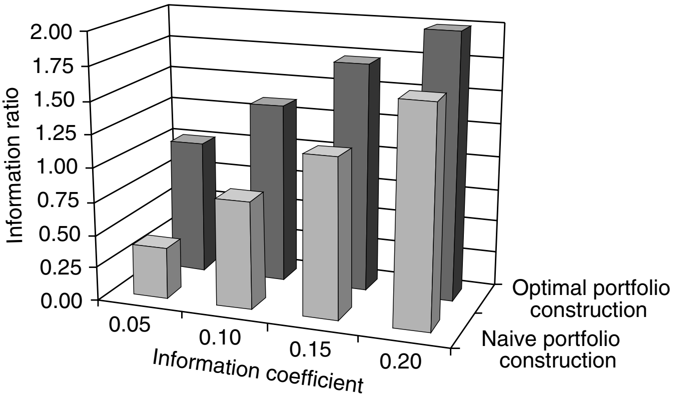
The data, obtained by Monte Carlo simulation, are returns expressed as basis points.

We would expect a manager with no skill to achieve a mean active return of zero. However, even a manager with no information could outperform the benchmark by about 100bp per year. Why should this be? Forecasting total returns and overweighting them each time they are positive would yield a structural long equity bias (overweight equities on average) as total returns can be expected to be positive in about 70% of all cases in our example. Being long equity 70% of the time would in turn allow the manager to collect parts of the equity risk premium. The effect is less pronounced for risk premiums (positive 60% of the time) and disappears for deviations from the fair equilibrium risk premium (positive deviations are as likely as negative deviations, ie, 50%). Structural biases should be avoided as they offer rewards for taking on structural risks – effectively, the collection of a risk premium – rather than for skill. In order to avoid a structural bet, there are two possibilities:

- Forecast deviations from the fair risk premium. This requires an estimate of the equilibrium risk premium. In the simplest case we take the CAPM (capital asset pricing model) and calculate $\pi_i = \beta_i \pi_{\text{m}}$, where π_{m} is the market risk premium and β_i reflects systematic risk.[23]
- Forecast whatever we like, but run the portfolio optimisation under a factor-neutrality constraint. Again, in a CAPM world this would require a portfolio beta of one with respect to the benchmark.

Both approaches should theoretically result in the same portfolios. In practice, however, they differ as this would require that we use

Figure 12.3 Information ratio and portfolio construction



Height of bars represents investment success for a given level of investment skill (skill measured as information coefficient).

exactly the same model for return refinement and risk decomposition. The example also shows that active return is a flawed concept. Essentially, it is a return figure with no adjustment for systematic risk, and it can easily be gamed. Adding a small to medium allocation of corporate bonds to a government bond fund will, in the long run, increase mean active return without any need for skill. As long as clients and consultants are unwilling to discourage this behaviour and euphemistically call it "style", it makes business sense for investment managers to engage in theses practices as it is a source of return the manger can bank on in the long run.[24]

The second question we wish to address in this section is whether portfolio construction brings any benefit to the investment process. Again we will use an example to illustrate the issues. Let us suppose we have monthly total return data for emerging market bonds (EMBI series) for Argentina, Brazil, Bulgaria, Ecuador, Mexico, Morocco, Nigeria, Panama, Peru, Poland, Russia and Venezuela covering the time period between January 1994 and January 2001.

We simulate an asset manager who tries to forecast cross-sectional emerging market returns for alternative levels of skill measured, as before, through the information coefficient. First we standardise the cross section of returns to make them standard normal, subtracting the cross-sectional mean, $\bar{R}_t$, and dividing by cross-sectional volatility, σ_t.[25] Next we generate cross-sectional forecasts

$$R^{\mathrm{f}}_{i,t} \sim \underbrace{\left(\mathrm{IC}\frac{R_i - \bar{R}_t}{\sigma_t} + \sqrt{1 - \mathrm{IC}^2}z\right)}_{\text{Bivariate standard normal}} \sigma_t + \bar{R}_t \tag{12.13}$$

where $R^{\mathrm{f}}_{i,t}$ denotes the simulated forecast for asset i in period t and z is a standard normal random variable for all assets and all time periods in our data sample. For each point in time we construct portfolios using two different portfolio construction methodologies:

1. *Optimal portfolio construction.* Use Equation 12.6 to construct TE-efficient portfolios that optimally trade off marginal risk against marginal return.
2. *Heuristic portfolio construction.* Overweight the top six countries and underweight the bottom six countries by equal amounts. Rankings are based on sorted forecast returns.

For the whole historical data set we generate a set of forecasts, calculate actual risk-adjusted performance (information ratio) for both portfolio construction methodologies and repeat history 1,000 times. The results are shown in Figure 12.3. For high information coefficients (an information coefficient of 0.2 is equivalent to an impressive R^2 of 0.44 in an out-of-sample backtest of a regression model), the choice of portfolio construction methodology makes little difference. This is no surprise as good return forecasts will dominate the portfolio construction process and risk management becomes increasingly less important for those who do have such forecasts. However, as the quality of the forecasts deteriorates, portfolio construction gains in importance. Although the level of information ratio falls for both methodologies, the spread increases for low to modest skill levels. Given that an information coefficient of 0.05 is regarded as good among active managers, it is not difficult to see why portfolio construction is considered important by institutional investors.

12.4 DUAL-BENCHMARK OPTIMISATION

So far we have assumed that investors have a single well-defined benchmark. Some investors, however, use more than one yardstick as a basis for measuring investment success. In most cases this practice is implicit rather than explicit – as in the case of a manager who is forced to manage relative risk but also needs to keep track of total risk to avoid blame for negative market returns as the plan sponsor will certainly not accept underperformance in a falling market.[26] Other examples are pension funds that have to follow peer benchmarks even though these are known to be inefficient in risk–return space, and plan sponsors who want to do well in terms of absolute return but at the same time want to beat a particular peer in their group.

Dual-benchmark optimisation requires finding the optimal portfolio weights, $\boldsymbol{w}_{\mathrm{p}}$, relative to two benchmarks $\boldsymbol{w}_{\mathrm{b}1}$ and $\boldsymbol{w}_{\mathrm{b}2}$, ie, aggregating (weighting) the two separate optimisation problems

$$\text{Utility}_i \approx \boldsymbol{w}'_{\mathrm{a}i}\boldsymbol{\mu} - \frac{1}{2\lambda_i}\boldsymbol{w}'_{\mathrm{a}i}\boldsymbol{\Omega}\boldsymbol{w}_{\mathrm{a}i}, \quad i = 1, 2 \tag{12.14}$$

in a coherent way. Note that we write the active position against the ith benchmark as $\boldsymbol{w}_{\mathrm{a}i} = \boldsymbol{w}_{\mathrm{p}} - \boldsymbol{w}_{\mathrm{b}i}$. If we set[27]

$$\theta = \left(\frac{1}{2\lambda_1}\right) \Big/ \left(\frac{1}{2\lambda_1} + \frac{1}{2\lambda_2}\right) \tag{12.15}$$

and use the weighting factor θ to generate

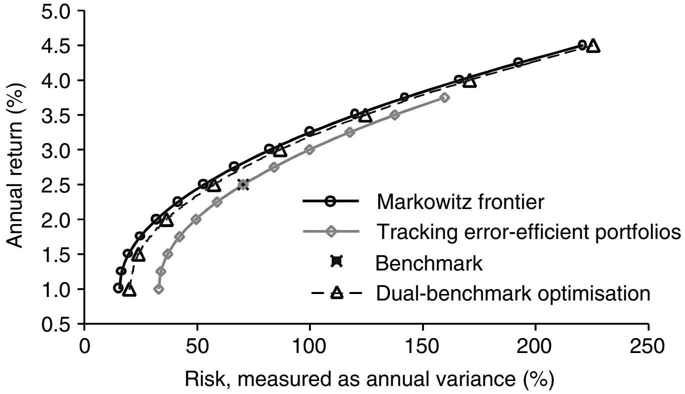
$$\left.\begin{aligned} \boldsymbol{w}_{\mathrm{b}} &= \theta\boldsymbol{w}_{\mathrm{b}1} + (1-\theta)\boldsymbol{w}_{\mathrm{b}2} \\ \frac{1}{2\lambda} &= \theta\frac{1}{2\lambda_1} + (1-\theta)\frac{1}{2\lambda_1} \end{aligned}\right\} \tag{12.16}$$

we can solve the traditional Markowitz problem $\text{Utility}_{\text{combined}} = \boldsymbol{w}'_{\mathrm{a}}\boldsymbol{\mu} - \frac{1}{2}\lambda\boldsymbol{w}'_{\mathrm{a}}\boldsymbol{\Omega}\boldsymbol{w}_{\mathrm{a}}$, where $\boldsymbol{w}_{\mathrm{a}} = \boldsymbol{w}_{\mathrm{p}} - \boldsymbol{w}_{\mathrm{b}}$. All that is required is to redefine the new single benchmark as a weighted average (with weights given in Equation 12.15) of the two previous benchmarks. A solution to this problem is given in Figure 12.4 for the example data set used in Section 12.2.

Moving away from the benchmark by introducing the objective of also beating cash – effectively the Markowitz objective – shifts the resulting frontier closer to the Markowitz frontier. The weightings will reflect the relative risk tolerance (importance) assigned to both objectives.[28] Any standard portfolio optimiser can be used for this. We do not have to worry about how to determine the risk tolerance parameter as we will vary this parameter anyway to trace out a new efficient frontier.

Alternatively, it has been suggested that the Pareto criterion can be applied to dual-benchmark optimisation.[29] This implies finding

Figure 12.4 Dual-benchmark optimisation and mean–variance efficiency



a solution such that it would be impossible to improve the risk-adjusted outperformance (utility) relative to benchmark one without losing utility from managing relative to benchmark two. Instead of weighting both utility functions we try to maximise, with respect to $\boldsymbol{w}_{\mathrm{p}}$

$$\begin{aligned}&\min(\text{Utility}_1, \text{Utility}_2)\\&\qquad\approx \min\left(\boldsymbol{w}'_{\mathrm{a1}}\boldsymbol{\mu} - \frac{1}{2\lambda_1}\boldsymbol{w}'_{\mathrm{a1}}\boldsymbol{\Omega}\boldsymbol{w}_{\mathrm{a1}}, \boldsymbol{w}'_{\mathrm{a2}}\boldsymbol{\mu} - \frac{1}{2\lambda_2}\boldsymbol{w}'_{\mathrm{a2}}\boldsymbol{\Omega}\boldsymbol{w}_{\mathrm{a2}}\right) \quad (12.17)\end{aligned}$$

For simplicity we assume that $\lambda = \lambda_1 = \lambda_2$.[30] In case $\lambda \to \infty$, we will try to maximise the minimum of the two active utilities. Equation 12.17 can be solved by maximising

$$\text{Utility}_1 = \boldsymbol{w}'_{\mathrm{a1}}\boldsymbol{\mu} - \frac{1}{2\lambda}\boldsymbol{w}'_{\mathrm{a1}}\boldsymbol{\Omega}\boldsymbol{w}_{\mathrm{a1}} \qquad (12.18)$$

adding a linear constraint to the usual optimisation problem

$$\underbrace{(\boldsymbol{w}_{\mathrm{b1}} - \boldsymbol{w}_{\mathrm{b2}})'\boldsymbol{\Omega}\boldsymbol{w}_{\mathrm{p}}}_{\text{Covariance}} = \underbrace{\lambda(\boldsymbol{w}_{\mathrm{b1}} - \boldsymbol{w}_{\mathrm{b2}})'\boldsymbol{\mu}}_{\text{Return difference}} + \underbrace{\tfrac{1}{2}(\boldsymbol{w}'_{\mathrm{b1}}\boldsymbol{\Omega}\boldsymbol{w}_{\mathrm{b1}} - \boldsymbol{w}'_{\mathrm{b2}}\boldsymbol{\Omega}\boldsymbol{w}_{\mathrm{b2}})}_{\text{Risk difference}}$$

If we focus on risks only ($\boldsymbol{\mu} = \mathbf{0}$), the linear constraint forces the risk difference between the two benchmarks to equal the covariance between the benchmark returns and a misfit portfolio that fills the difference between the two benchmarks.

In practice, dual-benchmark optimisation is of little, if any, consequence to practitioners as there is a clear separation of roles: consultants and sponsors set benchmarks (ie, total returns), while relative returns are controlled by investment managers. Although theoretically compelling, practitioners feel they have little guidance on how to set relative weights for competing objectives.

12.5 TRACKING ERROR AND ITS FUNDING ASSUMPTIONS

This section shows that changing the funding assumptions in calculating marginal contributions to tracking error can significantly enhance the interpretation and acceptance of portfolio risk decomposition and implied alpha calculation.[31] A meaningful decomposition of risk is a prerequisite for risk budgeting techniques, which are at the heart of modern investment processes.

12.5.1 Scalability

Ex ante TE is usually calculated using the well-known quadratic form of Equation 12.2.[32] In a world with no restrictions on short sales, benchmark and active positions can be perfectly separated. Active risk is generated by active positions, which can be overlaid on any benchmark. Each portfolio with a given TE (but possibly a different benchmark) has the same firm-specific alpha. Investment products differ only in their benchmark and level of aggressiveness. Varying levels of aggressiveness can be achieved through a linear scaling factor, ϕ, which leverages active positions according to the desired level of aggressiveness. A factor greater than unity gives a more aggressive portfolio than the model portfolio, and vice versa.[33] For example, doubling the scaling factor (doubling active positions) results in twice as much TE as before. The TE is also linear homogeneous in scaling factors and, hence, in active weights, as can be confirmed from the following equation

$$\begin{aligned}\sigma_{\mathrm{a}}\phi^{k} &= (\boldsymbol{w}_{\mathrm{a}}'\phi\boldsymbol{\Omega}\boldsymbol{w}_{\mathrm{a}}\phi)^{1/2}\\ &= (\phi^{2}\boldsymbol{w}_{\mathrm{a}}'\boldsymbol{\Omega}\boldsymbol{w}_{\mathrm{a}})^{1/2}\\ &= \phi(\boldsymbol{w}_{\mathrm{a}}'\boldsymbol{\Omega}\boldsymbol{w}_{\mathrm{a}})^{1/2}\\ &= \phi\sigma \qquad (12.19)\end{aligned}$$

This means that to create more aggressive portfolios one does not have to set up a new model portfolio. Scalability is a key requirement in modern investment processes as it allows new business to be accepted at low marginal costs and, hence, enables the typically high fixed costs to be quickly proportionalised.

12.5.2 Decomposition

It is well known in mathematics that linear homogeneous functions can be expressed using an Euler equation.[34] Applying this to the TE function yields the following result

$$\sigma_{\mathrm{a}} = \boldsymbol{w}'_{\mathrm{a}} \frac{\mathrm{d}\sigma}{\mathrm{d}\boldsymbol{w}_{\mathrm{a}}} = \sum_i w_{\mathrm{a}i} \frac{\mathrm{d}\sigma}{\mathrm{d}w_i}$$

Thus, TE can be decomposed into the summation of all marginal contributions to TE multiplied by active weights. Dividing both sides by TE yields the familiar result that all percentage contributions to risk (elasticity) have to sum to one. Again, a familiar property of linear homogeneous functions.

The implicit assumption in the above decomposition is that each position is funded from cash and, as cash has no volatility (at least not in a one-period framework), all risk comes from the asset positions. A problem of this assumption is that it does not reflect actual decision making. Portfolio managers do not fund their positions from cash but, rather, from other markets (pairs or triplets of positions form one trade) or from benchmarks (an overweight in one asset is financed by equally weighted underweights in all other assets). An overweight in the long end of the US fixed-income market is not likely to be funded from cash but rather from the one- to three-year US government sector (intramarket spread trade), the long end in Canada (intermarket spread trade), the long end of the US corporate sector (credit spread trade), or from multiple positions in medium- and short-maturity assets (barbell trade). However, it should be noted that although the TE decomposition will change, the total TE stays the same irrespective of the funding assumption.

12.5.2.1 Trade funding

Suppose we have a matrix, $\boldsymbol{T}$, of l active trades in k assets, each row of which represents one trade (giving the approach we are about to consider its alternative name of "trade funding", as opposed to cash funding). Positive entries, representing overweighted trade positions, are funded from negative (underweight) entries, as indicated below

$$\boldsymbol{T} = \begin{bmatrix} w_{\mathrm{a}1} & 0 & \cdots & -w_{\mathrm{a}k} \\ w_{\mathrm{a}2} & -w_{\mathrm{a}2} & -w_{\mathrm{a}3} & 0 \\ \vdots & & & \vdots \\ & & \cdots & \end{bmatrix}_{l\times k} \tag{12.20}$$

It can be assumed that the positions in each row sum to zero (imposing a restriction that does not, however, change the mathematics involved).[35] The aggregate portfolio position is the sum of the individual trade positions. Equipped with $\boldsymbol{T}$ we can now calculate the variance–covariance matrix of trades, $\boldsymbol{\Omega}_{\text{trades}}$, which is of dimension l by l

$$\boldsymbol{\Omega}_{\text{trades}} = \boldsymbol{T}'\boldsymbol{\Omega}\boldsymbol{T} \tag{12.21}$$

Numbers on the main diagonal of $\boldsymbol{\Omega}_{\text{trades}}$ represent the TE of the lth trade. Covariances between trades are captured by the off-diagonal entries. The TE of the total portfolio can be calculated by summing the individual entries of the new variance–covariance matrix of trades[36]

$$\sigma_{\text{a}} = (\mathbf{1}'\boldsymbol{\Omega}_{\text{trades}}\mathbf{1})^{1/2} \tag{12.22}$$

In order to calculate the marginal contribution to TE from each single trade we just take the first derivative with respect to **1**, defined to be an $l \times 1$ vector of 1s, as below

$$\frac{\mathrm{d}\sigma_{\text{a}}}{\mathrm{d}\mathbf{1}} = \boldsymbol{\Omega}_{\text{trades}}\mathbf{1}\sigma_a^{-1}, \qquad \sigma = \mathbf{1}\frac{\mathrm{d}\sigma_{\text{a}}}{\mathrm{d}\mathbf{1}} \tag{12.23}$$

TE decomposition follows the Euler equation-based approach introduced above. It is a straightforward matter to slice the variance–covariance matrix further into types of trades and to calculate the TEs arising from different decision areas

$$\boldsymbol{\Omega}_{\text{trades}} = \begin{bmatrix} \boldsymbol{\Omega}_{\text{DD}} & \boldsymbol{\Omega}_{\text{DC}} & \cdots & \boldsymbol{\Omega}_{\text{DCr}} \\ \vdots & \boldsymbol{\Omega}_{\text{CC}} & & \vdots \\ & & \ddots & \\ & & & \boldsymbol{\Omega}_{\text{CrCr}} \end{bmatrix} \tag{12.24}$$

Along the main diagonal we find the variance–covariance submatrixes for duration (D), currency (C) and credit (Cr), with the covariances between the different blocks of trade in off-diagonal positions.

Such decomposition becomes more important as TE budgeting is incorporated into investment processes.[37] If $\boldsymbol{T}$ becomes a diagonal matrix with active weights along the main diagonal and zero elsewhere, we get the same results as with the cash funding approach. Although this is certainly the most realistic approach, a risk management system using it will require additional information in the form of pairs, triplets and quadruplets of asset positions that form one trade. It becomes apparent that risk management tools have to reflect the respective investment process – which is why standard software suppliers do not offer risk decomposition by trades. Without knowledge of the portfolio manager's thoughts, any decomposition is impossible. Portfolios cannot simply be loaded from a fund accounting system into a risk management system as there is an infinity of trade decompositions that will deliver the same active weightings $\boldsymbol{w}'_{\mathrm{a}} = \mathbf{1}'\boldsymbol{T}$.

12.5.2.2 Benchmark funding

Portfolio managers often think of the benchmark return as their cost of funding.[38] We can express a long active position in asset i as being funded from $k-1$ equal underweights in the remaining assets. Benchmark funding is actually a special case of trade funding and can be captured by the following trade matrix

$$\boldsymbol{T}_{l\times l} = \begin{bmatrix} w_{\mathrm{a}1} & \dfrac{-w_{\mathrm{a}1}}{k-1} & \cdots & \dfrac{-w_{\mathrm{a}1}}{k-1} \\ w_{\mathrm{a}2} & \dfrac{-w_{\mathrm{a}2}}{k-1} & \cdots & \dfrac{-w_{\mathrm{a}2}}{k-1} \\ \vdots & & & \vdots \\ & & \cdots & \end{bmatrix} \tag{12.25}$$

The same mathematics applies as for the trade funding matrix in Equation 12.20.

One advantage of benchmark funding is that percentage contributions to risk are less often negative (which can, for example, happen if assets have similar risks and high correlations, as is the case for fixed-income markets). Since this is easier for portfolio managers to understand, acceptance of the procedure should rise. Percentage contributions to TE generally become negative if an active position and its marginal contribution to risk are of different sign. Suppose that asset i is underweight. As can be seen below, there is a direct effect (due to asset volatility) and an indirect effect (covariances)

$$\frac{\mathrm{d}\sigma_{\mathrm{a}}}{\mathrm{d}w_{\mathrm{a}i}} = \underbrace{2w_{\mathrm{a}i}\sigma_i^2}_{\text{Direct}} + \underbrace{2\sum_{i\neq j} w_{\mathrm{a}j}\sigma_{ij}}_{\text{Indirect}} \tag{12.26}$$

The direct effect on underweights will always be negative (negative active weight) as this implies that risk can be reduced by increasing the position towards zero. In most cases the marginal contribution for asset i will be negative unless there is, for example, a sizeable overweight in an asset with a high correlation with that asset. If the position in asset i is financed instead with $k - 1$ overweights in the remaining benchmark assets, the percentage allocation of TE for asset i remains positive as long as

$$\sigma_i^2 - \sum_{i \neq j} \frac{\sigma_{ij}}{k-1} > 0 \tag{12.27}$$

Expressed in words: the direct effect dominates the indirect effect provided that the variance is higher than the average covariance.[39] We can also use this approach to quantify correlation risk. Assuming equal correlation between asset i and all other assets j $(\rho_{ij} = \rho_i)$, we get, for benchmark funding

$$\frac{\mathrm{d}\sigma_{\mathrm{a}}^2}{\mathrm{d}\rho_i} = \frac{w_{\mathrm{a}i}^2 \sigma_i (\sum \sigma_j)}{k-1} \tag{12.28}$$

Correlation risk rises with bet size and asset volatility. However, the benchmark funding approach is not recommended as it leads to flawed decision making. This is because overweights do not have to beat the benchmark return. Positive alphas can also be generated by overweighting underperforming assets if the corresponding underweights are even stronger underperformers. Hence, the popular belief that it is difficult to outperform the benchmark if only a few assets are outperforming is, taken on its own, also not true.

12.5.2.3 Implied position analysis

For any active holding to be optimal the marginal contribution to risk and the marginal contribution to return (alpha) have to be equal to the information ratio.[40] This follows directly from optimality conditions. Improvements on the overall risk–return optimality could be achieved if the relationship between marginal risk and marginal return were not the same for all active positions. In the case of trade funding we get

$$\boldsymbol{\alpha} = \frac{\mathrm{d}\sigma_{\mathrm{a}}}{\mathrm{d}\mathbf{1}} \mathrm{IR} \tag{12.29}$$

where $\boldsymbol{\alpha}$ is the $k \times 1$ vector of active returns and IR is the information ratio. All active positions would plot on a straight line where the x-axis represents marginal risk and the y-axis represents marginal return. Portfolios which do not show this feature are either not optimised or are subject to constraints.

This concept can easily be applied to calculate implied asset movements. In the case of fixed-income management this could, for example, be used to calculate implied yield curve changes. Suppose that a positively weighted duration deviation (a weight-adjusted duration overweight) has been implemented. The return of this position can be approximated by

$$\alpha_i = -wdd_i \, \mathrm{d}y_i \tag{12.30}$$

Substituting this into the optimality condition for unconstrained optimal portfolios, we get

$$\mathrm{d}y_i = -\frac{\mathrm{IR}(\sum_{j=1}^{m} \zeta_{ij})/\sigma_{\mathrm{a}}}{wdd_i} \tag{12.31}$$

where ζ_{ij} refers to element i, j in the variance–covariance matrix of trades, $\boldsymbol{\Omega}_{\text{trades}}$. The marginal contribution to risk of the ith trade is calculated by summing the entries in the trade variance–covariance matrix along the ith row and dividing them by TE

$$\frac{\mathrm{d}\sigma_{\mathrm{a}}}{\mathrm{d}\mathbf{1}} = \frac{\mathrm{d}(\mathbf{1}'\boldsymbol{\Omega}_{\text{trades}}\mathbf{1})}{\mathrm{d}\mathbf{1}} = \frac{1}{2}\frac{\mathbf{2}\boldsymbol{\Omega}_{\text{trades}}\mathbf{1}}{\sigma_{\mathrm{a}}} = \begin{bmatrix} \dfrac{1}{\sigma_{\mathrm{a}}} \sum\limits_{j=1}^{m} \zeta_{1j} \\ \dfrac{1}{\sigma_{\mathrm{a}}} \sum\limits_{j=1}^{m} \zeta_{2j} \\ \dfrac{1}{\sigma_{\mathrm{a}}} \sum\limits_{j=1}^{m} \zeta_{mj} \end{bmatrix} \tag{12.32}$$

We are now able to calculate the implied market movements from a given set of active positions (nominal weights, weighted duration deviations, trade matrix, etc). These implied forecasts can then be compared with the manager's forecasts (if they exist) or his/her market assessment. If implied movements and expected market movements differ considerably, portfolio managers should rethink their positions and adjust them accordingly. This iterative process is also called "view optimisation". The main advantages of this technique are twofold. Traditional mean–variance-optimised portfolios often show (unrealistically) high information ratios. This does not apply to reverse optimisation as the information ratio is held constant and all returns are scaled accordingly. It also works as a reality check – whether the views implied by the current positions correspond with the forecast returns and investors' views.

We could also use the decomposition above to identify "best hedges", ie, to identify the risk-minimising trades.[41] Suppose we engage in l trades and each trade is leveraged up or down according to ϕ_j, $j = 1, \dots, l$. The leverage factor for, say, the fourth trade (ϕ_4) that minimises active portfolio risk (all things being equal) can be calculated from

$$\min\left(\sum_{i=1}^{l}\sum_{j=1}^{l}\phi_i\phi_j\zeta_{ij} + (\Delta\phi_4)^2\zeta_{44} + \sum_{i=1}^{l}\phi_i\Delta\phi_4\zeta_{i4}\right) \tag{12.33}$$

where Δ indicates the change in the respective leverage factor. The solution to Expression 12.33 gives us

$$\Delta\phi_4^* = \frac{\sum_{i=1}^{l}\phi_i\zeta_{i4}}{\zeta_{44}} \tag{12.34}$$

The right-hand side of Equation 12.34 contains the standardised covariance (beta) of the returns of the fourth trade with all other trades, including itself, which together make up active portfolio risk.

12.5.3 Summary

Tracking error is routinely used in modern asset management practice. At the same time risk budgeting techniques suggest that its decomposition (eg, what percentage of my TE comes from currency decisions?) is just as important as its size. In this section it has been shown that a more realistic decomposition of tracking error can enhance the understanding and acceptance of risk numbers. Employing this approach makes the thinking of portfolio managers more explicit and, thus, eases the task of portfolio construction with multiple benchmarks. Thinking of risk positions as trades allows portfolio managers to quickly monitor risk and the performance of trades.

12.6 TRADING BANDS FOR TACTICAL ASSET ALLOCATION

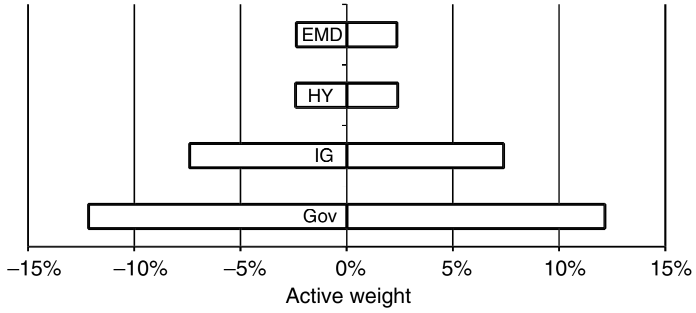
Trading bands are defined by the minimum and maximum allocations to a particular asset or asset class. Although they are a relic of a past in which TE calculations were not readily available, they are still required by consultants and plan sponsors. Their strength is the transparency they confer and the ease of legal action if broken; their weakness is their unreliability in achieving what they are supposed to do: control active risk. Trading bands that feel reasonable in times of high correlation will expose a fund to much higher active risk than planned if correlations break down. Conversely, rising correlations will reduce the dispersion of active returns around the benchmark but will also make it difficult to achieve the targeted active return for a given bet size (as defined by the trading bands). TE calculations are not immune to misestimation of active risks, but they will gradually pick up changes in the market environment. Either trading bands must be changed in line with changes in the marketplace (in which case they differ only philosophically from TE-based risk management) or they will fail to achieve their objective.

Every active fund can be written as the sum of all possible pairwise decisions. For example, in a fund containing four asset classes there are six pairwise asset allocation decisions: three decisions of the first asset against the other three assets, plus two decisions of the second asset against assets 3 and 4, plus one decision of asset 3 against asset 4. Although trading ranges mainly follow a risk objective, setting the ranges should also reflect the opportunity to add value. If an investment manager feels particularly confident about one of their decision pairs (for example, the balance decision in a balanced fund or a government bond/emerging market debt decision in a fixed-income product), this should be reflected in the trading bands.

Suppose we are developing the trading ranges for a global fixed-income product investing in government bonds and investment-grade, high-yield and emerging market debt. The development is a two-stage process. First we have to specify the maximum TE available for the tactical asset allocation decision layer. Assuming this to be 100bp, we then have to calculate the covariance matrix of active decisions by calculating the usual quadratic form using the trade matrix of all possible pairs. The next step is the most critical. We need to write down the information ratios for all pairwise decisions. It is assumed that the allocation decisions as between government bonds and all other asset classes are assigned an information ratio of 0.6, whereas it is considered that no skill is required to choose between high-yield and emerging market debt (EMD).

The process of actually deriving the optimal trading ranges is a quadratic optimisation problem in which optimal scaling factors ϕ_i (factors leveraging individual trade pairs) are derived and

Figure 12.5 Optimal trading bands



EMD, emerging market debt; HY, high-yield debt; IG, investment-grade debt; Gov, government bonds.

subsequently transformed into ranges

$$\alpha = \sum \mathrm{IR}_i \phi_i \zeta_i \quad \text{and} \quad \sigma_\mathrm{a} = \left(\sum_i \sum_j \phi_i \phi_j \zeta_{ij} \right)^{1/2} \tag{12.35}$$

After we have found the optimal scaling factors ϕ_i^*, we use them to leverage up the positions underlying the individual trades and find the optimal trading range for the ith asset from

$$w_j \in \left[\sum_i w_{\mathrm{b}i} + \phi_i^* \operatorname{abs}(w_{\mathrm{a}i}),\ w_{\mathrm{b}i} - \sum_i \phi_i^* \operatorname{abs}(w_{\mathrm{a}i}) \right] \tag{12.36}$$

For those who do not like theory, here is a practical solution to our example. It is apparent from Figure 12.5 that the more volatile decision areas (EMD, high-yield) are given tighter limits than the less volatile asset classes. This would only be reversed if one assumed unrealistically high information ratios for high-yield and emerging market debt. There are two main advantages of setting ranges as suggested above:

- All ranges are set simultaneously. Hence, any combination of active weights will yield a maximum TE of 100bp.
- The ranges reflect skill and risk.

This contrasts favourably with traditional techniques that define ranges in terms of risk only and limit the analysis to individual assets rather than looking at all ranges simultaneously.

APPENDIX A: GLOBAL FIXED-INCOME POLICY MODEL

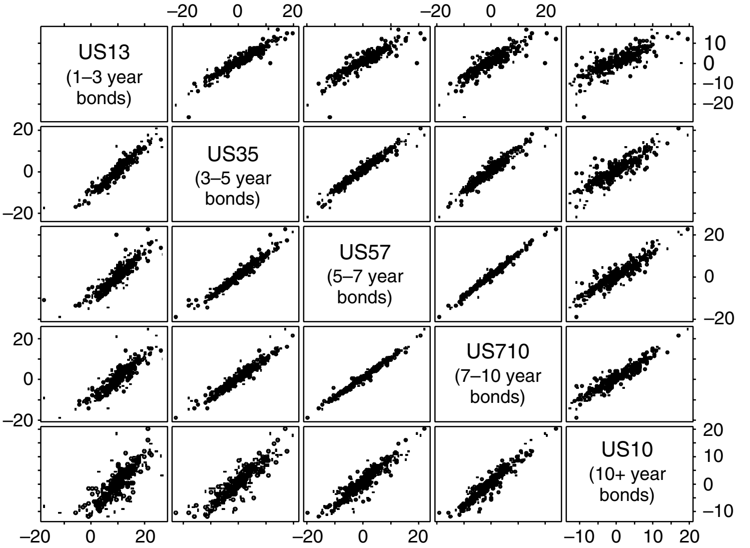
Basic issues in fixed-income risk management

Quantitative risk management, consistent forecasting and precise portfolio construction are central to any successful investment process. This appendix shows how to build an easily implemented fixed-income risk management and portfolio construction model for a benchmark-relative fixed-income investor.

Risk management models can only rarely be taken from the shelf. They have to be tailored to fit the specific underlying investment decision process. For volatility estimation it clearly makes a difference whether investment decisions are made frequently (volatility capture) or infrequently (structural alpha).[42] Whereas in the former process risk management has to capture changes in short-term volatility by using high-frequency data, in the latter it will rely more on structural relationships using low-frequency data. It has also to be decided whether the implementation of decisions is done on a "bucket" basis (the decision variable is the weighted duration deviation within a given bucket) or on an instrument by instrument basis (single bonds). As derivatives – the most efficient instrument for implementing portfolio allocation decisions – can deliver both linear and non-linear exposures, it should be decided whether portfolio construction will employ non-linear derivatives.

Our objective in this appendix is to model risk arising from macrobets taken on a bucket basis. Buckets are defined in terms of time to maturity (eg, all bonds with less than three years to maturity form the 1–3 year bucket). Typically, buckets are 1–3 years, 3–5 years, 5–7 years, 7–10 years and over 10 years. The view taken here is that yield curve movements are approximately parallel within buckets but less so between buckets. A duration-matched barbell (in which long- and short-maturity bonds are combined to match the duration of an intermediate-maturity bond) is not without interest rate risk as it certainly carries yield curve exposure. For benchmark-orientated investors this problem compounds as an active portfolio is equivalent to a leveraged investor.[43] The use of duration as a risk measure for international bond portfolios is even more flawed as it relies on the assumption of equal volatility and perfect correlation. Being overweight duration in Japan by one year against a one-year duration underweight in the US is certainly not a hedge against globally falling rates but is, rather, two separate duration positions.[44]

Figure 12.6 Daily yield curve changes for the US



Changes in bond yields across yield curve buckets are highly correlated within a given market, as can be seen from Figure 12.6. The bivariate scatter plots become cloudier as we move away from neighbouring buckets. This simply means that correlation is higher between the 1–3 year and 3–5 year buckets than between the 1–3 year and 10+ year buckets.

One of the best known statistical techniques for dealing with highly correlated series is principal component analysis. This is basically a data reduction technique that, in our case, tries to capture most of the volatility in the original series by creating a new, wholly artificial set of orthogonal series. For fixed-income markets the first three artificially created time series (the so-called principal components) explain most of the variance. The obvious advantage of this approach is that the number of parameters to be estimated in a risk model is greatly reduced. But can we give an economic meaning to these principal components? To answer this question we have to look at the sensitivities (loading) of yield curve changes to these principal components as in Figure 12.7.

The sensitivity of yields along the curve to changes in the first principal component is very similar. If the first principal component moves, all yields will move by approximately the same extent.

Figure 12.7 Ordered loading for the US

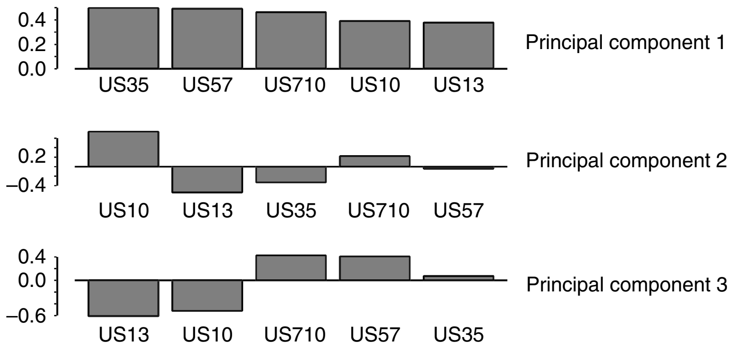
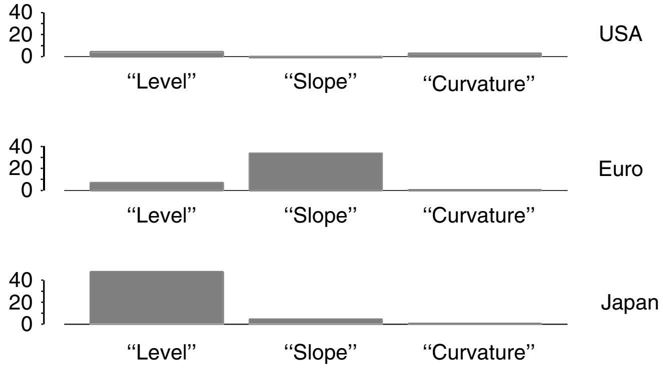


Figure 12.8 Decomposition of tracking error



This is called a "parallel yield curve move". If the second principal component moves upward, the short end yield will fall (negative loading), while long end yields will rise (positive loading).[45] This is called a steepening of the curve. If the third principal component rises, yields in the middle sector of the curve will rise while yields at the ends will fall. This is called a change in curvature.

Let us consider a global bond portfolio. We want to know what the TE decomposition will look like if we assume a duration-neutral barbell position in the US (long one-year-weighted duration deviation at the ends and short one year in the middle of the curve), a short-duration position in Japan (short 1.5 year in the 10+ bucket) and a duration-neutral yield curve position in Europe (long one year in the 10+ bucket and short one year at the short end).

The contributions to overall TE from the different active positions are plotted in Figure 12.8. The barbell position in the US makes the smallest contribution as the weighted duration deviations are small and the volatility of yield curve reshaping is also small. The small exposure to shifts in the level of yields (despite being duration-neutral) arises from factor loadings that are slightly different from each other. The spread position in Europe contributes most of its risk via the exposure to changes in the yield curve slope. The biggest exposure, however, arises from the duration exposure in Japan as the duration factor is more volatile than the others and because Japan shows little correlation with other developed bond markets.

Single-country model

At this stage we should become a little more formalistic in describing what is going on and how to replicate the suggested model. We start with decomposing the movements of the yield to maturity in the ith bucket of the jth country, Δy_i^j, into movements of $n_f = 3$ factors plus noise (unexplained variation)

$$\underbrace{\Delta y_i^j}_{\substack{\text{Yield}\\ \text{change}}} = \underbrace{\sum_{k=1}^{n_f} a_{i,k}^j \Delta f_k^j}_{\substack{\text{Sensitivity-weighted}\\ \text{factor changes}}} + \underbrace{\varepsilon_i^j}_{\text{Noise}} \tag{12.37}$$

As we have seen in Figure 12.7, each factor can be interpreted as a change in the level, slope or curvature of the yield curve.[46] The loadings $a_{i,k}^j$ represent the sensitivity of the ith bucket in the jth country to changes in the kth country-specific factor.

Assume further that we model $n_b = 5$ yield curve buckets in n_c countries. The factors and loadings for each country are simultaneously extracted by using principal component analysis on the exponentially weighted covariance matrix of yield curve changes. The global variance–covariance matrix of principal components is then constructed from the time series of factor scores. The correlation structure of international bond markets can also be captured with principal component analysis. For example, the first components in the US and Canada are highly correlated but show only little correlation with Japan. Therefore, by employing principal component analysis we can not only reduce the number of parameters to be estimated but also put economic interpretations on the factors driving yield curve changes that are familiar to portfolio managers. This enhances both the power of the model and its acceptability to portfolio managers.

How can we translate yield change risk into return risk? It is well known that total returns for a given yield curve bucket can be approximated by

$$R_i^j \approx \underbrace{y_i^j}_{\text{Carry}} - \underbrace{D_i^j \Delta y_i^j}_{\substack{\text{Effect of}\\\text{duration}}} + \underbrace{\tfrac{1}{2} C_i^j (\Delta y_i^j)^2}_{\substack{\text{Convexity}\\\text{adjustment}}} \tag{12.38}$$

Applying the variance operator we get

$$\operatorname{var}(R_i^j) \approx (D_i^j)^2 \operatorname{var}(\Delta y_i^j) \tag{12.39}$$

given the observation that yield and convexity effects are likely to be small. From principal component analysis we know that[47]

$$\operatorname{var}(\Delta y_i^j) = \sum_{k=1}^{n_f} (a_{i,k}^j)^2 \operatorname{var}(\Delta f_k^j) + \operatorname{var}(\varepsilon_i^j) \tag{12.40}$$

as the factor changes are independent by construction and not correlated with the error term. Taking the country-specific loading matrix and the variance–covariance matrix of factor movements, we get the covariance matrix of yield changes, $\boldsymbol{V}_j$

$$\boldsymbol{V}_j = \boldsymbol{A}_j \boldsymbol{\Phi}_j \boldsymbol{A}_j' + \boldsymbol{\Theta}_j \tag{12.41}$$

where $\boldsymbol{A}_j$ and $\boldsymbol{\Theta}_j$ are, respectively, the matrixes of loadings and of unexplained variations. The full matrixes in Equation 12.41 are given below

$$\boldsymbol{V}_j = \begin{bmatrix} \operatorname{var}(\Delta y_1^j) & \operatorname{cov}(\Delta y_1^j, \Delta y_2^j) & \cdots & \operatorname{cov}(\Delta y_1^j, \Delta y_5^j) \\ \operatorname{cov}(\Delta y_2^j, \Delta y_1^j) & \operatorname{var}(\Delta y_2^j) & & \\ \vdots & & \ddots & \\ \operatorname{cov}(\Delta y_5^j, \Delta y_1^j) & & & \operatorname{var}(\Delta y_5^j) \end{bmatrix}_{n_b \times n_b}$$

$$\boldsymbol{A}_j = \begin{bmatrix} a_{11}^j & a_{12}^j & a_{13}^j \\ a_{21}^j & a_{22}^j & a_{23}^j \\ \vdots & & \\ a_{51}^j & a_{52}^j & a_{53}^j \end{bmatrix}_{n_b \times n_f}$$

$$\boldsymbol{\Phi}_j = \begin{bmatrix} \mathrm{var}(\Delta f_1^j) & 0 & 0 \\ 0 & \mathrm{var}(\Delta f_2^j) & 0 \\ 0 & 0 & \mathrm{var}(\Delta f_3^j) \end{bmatrix}_{n_f \times n_f}$$

$$\boldsymbol{\Theta}_j = \begin{bmatrix} \mathrm{var}(\varepsilon_1^j) & 0 & \cdots & 0 \\ 0 & \mathrm{var}(\varepsilon_2^j) & & \\ \vdots & & \ddots & \\ 0 & & & \mathrm{var}(\varepsilon_5^j) \end{bmatrix}_{n_b \times n_b}$$

The variance of changes in the ith yield bucket not explained by the first three principal components is denoted as $\mathrm{var}(\varepsilon_i^j)$. From the variance–covariance matrix of yield changes it is only a small step to that of total returns

$$\begin{aligned} \boldsymbol{\Omega}_j &= \underbrace{\boldsymbol{D}_j\boldsymbol{A}_j\boldsymbol{\Phi}_j\boldsymbol{A}_j'\boldsymbol{D}_j'}_{\substack{\text{Variance–covariance}\\ \text{factors}}} + \underbrace{\boldsymbol{D}_j\boldsymbol{\Theta}_j\boldsymbol{D}_j'}_{\substack{\text{Variance–covariance}\\ \text{yield changes}}} \\ &= \underbrace{\boldsymbol{D}_j\boldsymbol{V}_j\boldsymbol{D}_j'}_{\substack{\text{Variance–covariance}\\ \text{total returns}}} \end{aligned} \tag{12.42}$$

where

$$\boldsymbol{\Omega}_j = \begin{bmatrix} \mathrm{var}(R_1^j) & \mathrm{cov}(R_1^j, R_2^j) & \cdots & \mathrm{cov}(R_1^j, R_5^j) \\ \mathrm{cov}(R_2^j, R_1^j) & \mathrm{var}(R_2^j) & & \\ \vdots & & \ddots & \\ \mathrm{cov}(R_5^j, R_1^j) & & & \mathrm{var}(R_5^j) \end{bmatrix}_{n_b \times n_b}$$

$$\boldsymbol{D}_j = \begin{bmatrix} D_1^j & 0 & \cdots & 0 \\ 0 & D_2^j & & \\ \vdots & & \ddots & \\ 0 & & & D_5^j \end{bmatrix}_{n_b \times n_b}$$

and D_i^j denotes the duration of the ith bond bucket (the 1–3 year bucket if $i = 1$) for the jth country. As most of the variance is explained by the first three principal components, the second term in Equation 12.42, $\boldsymbol{D}_j\boldsymbol{\Theta}_j\boldsymbol{D}_j'$, is often dropped. For ease of exposition we follow this convention here. The reader should, however, be warned that this assumption makes a significant difference for some countries – notably Japan and Switzerland. In order to calculate the portfolio total volatility we have to apply the vector of portfolio weights to the truncated Equation 12.42

$$\left.\begin{aligned} \sigma_{\text{total}} &= (\boldsymbol{w}_j' \boldsymbol{D}_j \boldsymbol{A}_j \boldsymbol{\Phi}_j \boldsymbol{A}_j' \boldsymbol{D}_j' \boldsymbol{w}_j)^{1/2} \\ \boldsymbol{w}_j' &= [\, w_1^j \quad w_2^j \quad \Lambda \quad w_5^j \,] \end{aligned}\right\} \tag{12.43}$$

Alternatively, we can write this using weighted duration (portfolio weight times portfolio duration)

$$\boldsymbol{wd}_j' = \boldsymbol{w}_j' \boldsymbol{D}_j = [\, w_1^j D_1^j \quad w_2^j D_2^j \quad \Lambda \quad w_5^j D_5^j \,]_{1\times n_b}$$

and apply this to the variance–covariance matrix of yield changes

$$\begin{aligned} \sigma &= (\boldsymbol{wd}_j' \boldsymbol{A}_j \boldsymbol{\Phi}_j \boldsymbol{A}_j' \boldsymbol{wd}_j)^{1/2} \\ &= (\boldsymbol{wd}_j' \boldsymbol{V}_j \boldsymbol{wd}_j)^{1/2} \end{aligned} \tag{12.44}$$

TE can now be easily calculated by using the vector of weighted duration deviations, $\boldsymbol{wdd}$. Weighted duration deviation is calculated by multiplying the weight with which a portfolio is invested into the ith bucket by the portfolio duration in that bucket and subtracting its benchmark equivalent. Effectively, it is the weighted portfolio duration minus the weighted benchmark duration

$$\boldsymbol{wdd}_j' = [\, w_1^j D_1^j - w_{1,\text{B}}^j D_{1,\text{B}}^j \quad \cdots \quad w_5^j D_5^j - w_{5,\text{B}}^j D_{5,\text{B}}^j \,]_{1\times n_b} \tag{12.45}$$

Again, TE is given by the usual quadratic form

$$\begin{aligned} \sigma_{\text{active}} &= (\boldsymbol{wdd}_j' \boldsymbol{A}_j \boldsymbol{\Phi}_j \boldsymbol{A}_j' \boldsymbol{wdd}_j)^{1/2} \\ &= (\boldsymbol{wdd}_j' \boldsymbol{V}_j \boldsymbol{wdd}_j)^{1/2} \end{aligned} \tag{12.46}$$

Alternatively, the TE can be calculated using duration-adjusted weights

$$\omega_i^j = w_{\text{adj},i}^j - w_{i,\text{B}}^j = w_i^j \frac{D_i^j}{D_{i,\text{B}}^j} - w_{i,\text{B}}^j = \frac{wdd_i^j}{D_{i,\text{B}}^j} \tag{12.47}$$

and applying them to the variance–covariance matrix of total returns

$$\left.\begin{aligned} \sigma_{\text{a}} &= (\omega_j^{\text{T}} \boldsymbol{\Omega}_j \omega_j)^{1/2} \\ \omega_j' &= [\, \omega_1^j \quad \omega_2^j \quad \Lambda \quad \omega_5^j \,] \end{aligned}\right\} \tag{12.48}$$

MULTI-COUNTRY MODEL

How does the preceding model change in a multi-country setting? Very little, as we shall now see. In principle, Equation 12.41 remains much the same

$$\boldsymbol{V}^* = \boldsymbol{A}^*\boldsymbol{\Phi}^*\boldsymbol{A}^{*\prime} + \boldsymbol{\Theta}^* \tag{12.49}$$

where

$$\boldsymbol{A}^* = \begin{bmatrix} A_1 & 0 & & \\ 0 & A_2 & & \\ & & \ddots & \\ & & & \boldsymbol{A}_{n_c} \end{bmatrix}_{(n_c n_b)\times(n_c n_b)}$$

$$\boldsymbol{\Phi}^* = \begin{bmatrix} \Phi_{11} & \Phi_{12} & \cdots & \\ \Phi_{21} & \Phi_{22} & & \\ \vdots & & \ddots & \\ & & & \Phi_{n_c n_c} \end{bmatrix}_{(n_f n_c)\times(n_f n_c)}$$

$$\boldsymbol{\Theta}^* = \begin{bmatrix} \theta_1 & 0 & \cdots & 0 \\ 0 & \theta_2 & & \\ \vdots & & \ddots & \\ 0 & & & \theta_{n_c} \end{bmatrix}_{(n_c n_b)\times(n_c n_b)}$$

It follows directly that $\boldsymbol{\Omega}^* = \boldsymbol{D}^*\boldsymbol{A}^*\boldsymbol{\Phi}^*\boldsymbol{A}^{*\prime}\boldsymbol{D}^{*\prime} + \boldsymbol{D}^*\boldsymbol{\Theta}^*\boldsymbol{D}^{*\prime}$, where

$$\boldsymbol{D}^* = \begin{bmatrix} D_1 & 0 & \cdots & 0 \\ 0 & D_2 & & \\ \vdots & & \ddots & \\ 0 & & & D_{n_c} \end{bmatrix}_{(n_c n_b)\times(n_c n_b)}$$

All other calculations follow accordingly. Instead of using the local loading, factor correlation, weight and duration, we use their global counterparts – here labelled with a star. For example

$$\begin{aligned} \sigma_{\text{active}} &= (\boldsymbol{wdd}^{*\prime}\boldsymbol{A}^*\boldsymbol{\Omega}^*\boldsymbol{A}^{*\prime}\boldsymbol{wdd}^*)^{1/2} \\ &= (\boldsymbol{wdd}^{*\prime}\boldsymbol{V}^*\boldsymbol{wdd}^*)^{1/2} \end{aligned} \tag{12.50}$$

INCLUDING CREDIT

A simple way of including corporate bonds is to model the investment-grade corporate bond market as a whole, not distinguishing between AAA and BBB ratings.[48] The corporate bond spread volatility arising from different maturity buckets is modelled directly as we recognise that movements in corporate bond yields, Δyc_i^j, can be decomposed into changes in government bond yields, Δy_i^j, and changes in the credit spread, Δs_i^j

$$\begin{aligned}\Delta yc_i^j &= \Delta y_i^j + \Delta s_i^j \\ &= \sum_{k=1}^{K} a_{i,k}^j \Delta f_k^j + \sum_{k=1}^{K} \underbrace{b_{i,k}^j}_{\substack{\text{Spread}\\\text{factor}\\\text{loading}}} \underbrace{\Delta fc_k^j}_{\substack{\text{Spread}\\\text{factor}}} + e_i^j \end{aligned} \tag{12.51}$$

Changes in credit spread are modelled using principal components, where the components can be interpreted as changes in level, slope and curvature of the credit curve with loading, $b_{i,k}^j$. For a one-country model including credit (along the five yield curve buckets), the matrixes involved change as outlined below

$$\boldsymbol{V}_j = \boldsymbol{A}_j \boldsymbol{\Phi}_j \boldsymbol{A}_j' + \boldsymbol{\Theta}_j \tag{12.52}$$

where the covariance matrix of yield changes can be decomposed into

$$\boldsymbol{V}_j = \begin{bmatrix} \boldsymbol{V}_j^{11} & \boldsymbol{V}_j^{12} \\ \boldsymbol{V}_j^{21} & \boldsymbol{V}_j^{22} \end{bmatrix}_{(2n_b)\times(2n_b)}$$

$$\boldsymbol{V}_j^{11} = \begin{bmatrix} \operatorname{var}(\Delta y_1^j) & \operatorname{cov}(\Delta y_1^j, \Delta y_2^j) & \cdots \\ \vdots & \ddots & \\ \operatorname{cov}(\Delta y_5^j, \Delta y_1^j) & & \operatorname{var}(\Delta y_5^j) \end{bmatrix}_{n_b\times n_b}$$

$$\boldsymbol{V}_j^{12} = \begin{bmatrix} \operatorname{cov}(\Delta y_1^j, \Delta yc_1^j) & \cdots & \operatorname{cov}(\Delta y_1^j, \Delta yc_5^j) \\ & \ddots & \\ & & \end{bmatrix}_{n_b\times n_b}$$

$$\boldsymbol{V}_j^{22} = \begin{bmatrix} \operatorname{var}(\Delta yc_1^j) & \cdots & \operatorname{cov}(\Delta yc_1^j, \Delta yc_5^j) \\ & \ddots & \\ & & \operatorname{var}(\Delta yc_5^j) \end{bmatrix}_{n_b\times n_b}$$

and the new matrix of factor loadings (yield curve and credit) is given as

$$\boldsymbol{A}_j = \left[\begin{array}{ccc|ccc}
a^j_{11} & a^j_{12} & a^j_{13} & 0 & 0 & 0 \\
a^j_{21} & a^j_{22} & a^j_{23} & 0 & 0 & 0 \\
\vdots & & & \vdots & & \\
\vdots & & & \vdots & & \\
a^j_{51} & a^j_{52} & a^j_{53} & 0 & 0 & 0 \\
\hline
a^j_{11} & a^j_{12} & a^j_{13} & b^j_{11} & b^j_{12} & b^j_{13} \\
 & & & b^j_{21} & b^j_{22} & b^j_{23} \\
\vdots & & & \vdots & & \\
\vdots & & & \vdots & & \\
a^j_{51} & a^j_{52} & a^j_{53} & b^j_{51} & b^j_{52} & b^j_{53}
\end{array}\right]_{2n_b \times 2 \times n_f}$$

where the different parts of the matrix are, top left, the yield curve risk of government bonds, top right, the zero spread risk of government bonds, bottom left, government risk in corporates, and bottom right, spread factor risk. The combined matrix of factor volatilities is given as

$$\boldsymbol{\Phi}_j = \left[\begin{array}{ccc|ccc}
\operatorname{var}(\Delta f^j_1) & 0 & 0 & \operatorname{cov}(\Delta f^j_1, \Delta fc^j_1) & \cdots & \\
0 & \operatorname{var}(\Delta f^j_2) & 0 & \vdots & \ddots & \\
0 & 0 & \operatorname{var}(\Delta f^j_3) & & & \\
\hline
\operatorname{cov}(\Delta f^j_1, \Delta fc^j_1) & \cdots & & \operatorname{var}(\Delta fc^j_1) & 0 & 0 \\
\vdots & \ddots & & 0 & \operatorname{var}(\Delta fc^j_2) & 0 \\
 & & & 0 & 0 & \operatorname{var}(\Delta fc^j_3)
\end{array}\right]_{2n_f \times 2n_f}$$

Here the parts are government factors (top left), factor covariation (top right and bottom left) and spread factors (bottom right).

Apart from knowing the size of the TE taken it is vital for portfolio managers to know which positions are more diversifying than others and how much TE can be attributed to a specific factor, country or bucket

$$\frac{\mathrm{d}\sigma_{\text{active}}}{\mathrm{d}\boldsymbol{wdd}} = \frac{\boldsymbol{V}^* \boldsymbol{wdd}^*}{\sigma_{\text{active}}}$$

$$\frac{\mathrm{d}\sigma_{\text{active}}}{\mathrm{d}\boldsymbol{f}} = \frac{\boldsymbol{\Omega}^* \boldsymbol{A}^{*\prime} \boldsymbol{wdd}^*}{\sigma_{\text{active}}}$$

Calculating the percentage contributions from here is a straightforward exercise. This approach can easily be overlaid with trade funding.

EXERCISE

1. Replicate Figure 12.1.

[1] See Sharpe (1981).

[2] See Roll (1992).

[3] Alternatively, mean absolute deviation (MAD) has also been labelled as TE as in Rudolf *et al* (1999)

$$\sigma_{\text{MAD}} = \frac{1}{T-1}\sum_{t=1}^{T} |r_{\text{a},t} - \bar{r}_{\text{a}}|$$

who claim that linear fee arrangements would make MAD better suited to capturing risk than Equation 12.1. We will ignore the latter definition because it has little impact on practitioners and also because the argument is not always convincing. Even if the fee arrangement is linear, the relationship between performance and fees will be non-linear as better performance leads to more assets under management and vice versa.

[4] Alternatively, assume that we deal with log active returns.

[5] The technology stock bubble of 2000 is a good example of this. Many fund managers engaged in very "sticky" bets (so termed because fund managers held them for a long time) while at the same time the trend in markets made historical volatility small. This led to a marked underestimation of weights-based TE versus return-based TE as the latter increased sharply upwards due to increased volatility and reduced correlation between stocks.

[6] For a review of variance ratio tests, see Campbell *et al* (1997).

[7] See Satchell and MacQueen (1998).

[8] For this reason some index trackers do not minimise TE but instead minimise transaction cost (return component) with a TE constraint.

[9] Assuming that normality is another caveat when trying to link TE to performance. However, non-normality might be less of a problem for diversified active positions as a large number of over- and underweights would bring the Central Limit Theorem into play. Skewness might diversify away much quicker than in long-only portfolios as the negative skewness of individual asset returns is transformed into a positive skewness for underweights and, hence, over- and underweights quickly reduce portfolio skewness.

[10] See Gardner *et al* (2000).

[11] See Fishwick (1999).

[12] See Scowcroft and Sefton (2001).

[13] See Michaud (1989).

[14] See Muller (1993).

[15] See Satchell and Hwang (2001).

[16] Suppose we can express uncertainty in weights as $\boldsymbol{w}_{\text{a}} = \bar{\boldsymbol{w}}_{\text{a}} + \boldsymbol{v}$, where $\boldsymbol{w}_{\text{a}} \approx N(\mathbf{0}, \boldsymbol{\Sigma})$. We can then decompose the squared TE into $\sigma^2 = \bar{\boldsymbol{w}}_{\text{a}}\boldsymbol{\Omega}\bar{\boldsymbol{w}}_{\text{a}} + \boldsymbol{\mu}_{\text{a}}\boldsymbol{\Omega}\boldsymbol{\mu}_{\text{a}} + \text{tr}(\boldsymbol{\Omega\Sigma})$. As the last two terms are always positive (positive definite matrix and trace of a covariance matrix), this will always be higher than the conventional calculation.

[17] See Hartmann (2002) versus Lawton (2001).

[18] See Roll (1992) or Wilcox (1994, 2000).

[19] This is only true as long as we measure risk as variance, as only variance is additive.

[20] See Leupold (1996, pp. 99–100).

[21] In practice, active positions and benchmark are much more difficult to separate as binding long-only constraints prohibit the effective use of short sales. As a reaction, asset managers cluster their books of business around a few model portfolios and in general become increasingly reluctant to accept non-standard ("odd" benchmark) business.

[22] More on different aspects of the information coefficient can be found in Grinold and Kahn (2000).

[23] The equilibrium risk premium is most likely not stable. Hence, Lee (2000) suggests using a filter mechanism to take time variation into account.

[24] Instead of defining $\bar{\alpha} = \bar{R}_i - \bar{R}_b$, we should estimate $\bar{\alpha} = (R_i - c) - \hat{\beta}(R_b - c)$. This would resolve the problem in a CAPM world.

[25] Cross-sectional volatility is a function of correlation and volatility. It is, however, not true that a higher cross section of returns leads to more attractive risk/return opportunities for active management, ie, a higher information ratio. Even if assets were highly correlated or volatility were low, we could always leverage up all active positions to generate the appropriate active return (assuming leverage were possible). The information ratio stays constant and is a function of skill (information coefficient) and effort (number of bets) only.

[26] See Kritzman and Rich (1998).

[27] See Wang (1999).

[28] The case $\theta = \frac{1}{2}$ does not mean that both benchmarks are tracked equally.

[29] The Pareto solution was introduced by Shectman (2000).

[30] General optimisers will find solutions to deviating cases also. We are, however, concerned with solutions that can be handled within the "normal" mean–variance framework.

[31] See Scherer (2001).

[32] Although the TE computation is similar to the calculation of portfolio volatility, it reacts differently to changes in market environment. Low-correlation assets reduce total risk but increase TE as active positions can drift apart more freely. Hence, rising correlations will tend to make long short positions better hedges, which will result in a drop in TE (unless overcompensated by rising volatility).

[33] Scaling factors are restricted to being positive as negative numbers would result in a change in positions (overweights become underweights and vice versa).

[34] See Berck and Sydsaeter (1991, p. 15, Equation (3.17)).

[35] Active nominal weights have to sum to zero for portfolios which cannot use leverage. However, for fixed-income portfolios it makes more sense to describe active positions in terms of weighted duration deviations. These do not have to sum to one.

[36] We could even generalise further at this stage and give weights different from unity to individual trades.

[37] See Scherer (2000).

[38] While this is not correct – as it is well known that active positions are independent of benchmarks (this argument is weakened for high-TE products or narrow benchmarks) – it still appears to be widespread.

[39] Inserting $w_{\mathrm{a}j} = w_{\mathrm{a}i}/(k-1)$ into the definition of the marginal contribution to risk yields the result in the text.

[40] The information ratio shows the relationship between excess return and TE. The higher the information ratio, the more favourable this trade-off becomes. However, very high information ratios are hardly credible as an information ratio of 1.96 implies only a 2.5% probability of underperforming with respect to a given benchmark.

[41] Litterman and Winkelmann (1996).

42 Structural alphas arise from a systematic bias in portfolio construction. These biases could arise from being structurally long equities in a balanced fund, constantly overweighting credit or constantly underweighting Japan. Along this structural position the investor tries to capture the volatility in the according asset. The higher the volatility, the longer the distance the investor can capture and the higher the potential out (under)performance from volatility capture. See Lee (2000).

43 An active portfolio is effectively a zero-investment long–short portfolio.

44 Statistical analysis shows that a global interest rate factor explains only about 50% of the variance in global yield changes as opposed to about 90% for national yield changes.

45 It should be noted that the exposures are ordered by absolute size.

46 Instead of using principal component analysis – which has the disadvantage that the principal components are orthogonal by construction – we might want to define the corresponding yield curve movements. Define yield curve factors according to

$$\begin{aligned}
\Delta l &= \text{Level}_t - \text{Level}_{t-1} \\
&= \frac{1}{5}\sum_i y^j_{i,t} - \frac{1}{5}\sum_i y^j_{i,t-1} \\
\Delta s &= \text{Slope}_t - \text{Slope}_{t-1} \\
&= (y^j_{5,t} + y^j_{1,t}) - (y^j_{5,t-1} - y^j_{1,t-1}) \\
\Delta c &= \text{Curvature}_t - \text{Curvature}_{t-1} \\
&= (y^j_{5,t} + y^j_{1,t} - 2y^j_{2,t}) - (y^j_{5,t-1} + y^j_{1,t-1} - 2y^j_{2,t-1})
\end{aligned}$$

and run OLS regression to calculate the specific loading (exposure)

$$\Delta y^j_i = \alpha^j_i + \alpha^j_{i,l}\Delta l^j + \alpha^j_{i,s}\Delta s^j + \alpha^j_{i,c}\Delta c^j + \varepsilon^j_i$$

47 For an application of principal component analysis to fixed-income markets, see Golub and Tilman (2000).

48 Rather than modelling the credit spread using principal component analysis, we might want to try swap spreads

$$\Delta s^j_{i,\text{rating}} = \beta_0 + \beta_1 \Delta s^j_{\text{swap}} + \upsilon^j_{i,\text{rating}}$$

However, the number of parameters being estimated here may be too high to ensure sufficient accuracy in the results.

REFERENCES

Berck, P., and K. Sydsaeter, 1991, *Economists Mathematical Manual*, Second Edition (Berlin: Springer).

Campbell, J., A. Lo and A. MacKinlay, 1997, *The Econometrics of Financial Markets* (Princeton, NJ: Princeton University Press).

Fishwick, E., 1999, "Unexpectedly Large or Frequent Extreme Returns in Active TE Portfolios", Franklin Portfolio Associates.

Gardner, D., D. Bowie, M. Brooks and M. Cumberworth, 2000, "Predicted TEs: Fact or Fantasy", Working Paper, Faculty and Institute of Actuaries.

Golub, B., and M. Tilman, 2000, *Risk Management* (New York: John Wiley & Sons).

Grinold, R., and R. Kahn, 2000, *Active Portfolio Management*, Second Edition (New York: McGraw-Hill).

Hartmann, S., 2002, "Laying the Foundations", ABN AMRO Research Paper, January.

Kritzman, M., and D. Rich, 1998, "Risk Containment for Investors with Multivariate Utility Functions", *Journal of Derivatives* 5, pp. 178–470.

Lawton, C., 2001, "An Alternative Calculation of TE", *Journal of Asset Management* 2, pp. 223–34.

Lee, W., 2000, *Theory and Methodology of Tactical Asset Allocation*, Frank J. Fabozzi Series (New York: John Wiley & Sons).

Leupold, T., 1996, *Benchmarkorientierte Portfoliooptimierung* (Bern: Haupt).

Litterman, R., and K. Winkelmann, 1996, "Managing Market Exposure", *Journal of Portfolio Management* 22, pp. 32–49.

Michaud, R., 1989, "The Markowitz Optimization Enigma: Is Optimized Optimal?", *Financial Analysts Journal* 45, pp. 31–42.

Muller, P., 1993, "Empirical Tests of Biases in Equity Portfolio Optimization", in S. Zenios (ed), *Financial Optimization* (Cambridge University Press).

Roll, R., 1992, "A Mean Variance Analysis of TE", *Journal of Portfolio Management* 18, pp. 13–22.

Rudolf, M., H. Wolter and H. Zimmermann, 1999, "A Linear Model for Tracking Error Minimisation", *Journal of Banking and Finance* 23, pp. 85–103.

Satchell, S., and S. Hwang, 2001, "TE: *Ex Ante* Versus *Ex Post* Measures", *Journal of Asset Management* 2, pp. 241–6.

Satchell, S., and J. MacQueen, 1998, "Why Forecast TE Seem Sometimes Inconsistent with Actual Performance", Working Paper, Alpha Strategies Ltd.

Scherer, B., 2000, "Preparing the Best Risk Budget", *Risk* 13, pp. 3–32.

Scherer, B., 2001, "A Note on TE Funding Assumptions", *Journal of Asset Management* 2, pp. 235–40.

Scowcroft, A., and J. Sefton, 2001, "Do TEs Reliably Estimate Portfolio Risk?", *Journal of Asset Management* 2, pp. 205–22.

Sharpe, W., 1981, "Decentralised Investment Management", *Journal of Finance* 36, pp. 217–34.

Shectman, P., 2000, "Multiple Benchmarks and Multiple Sources of Risk", Working Paper, Northfields.

Wang, M., 1999, "Multiple-Benchmark and Multiple Portfolio Optimization", *Financial Analysts Journal* 55, pp. 63–72.

Wilcox, J., 1994, "EAFE is for Wimps", *Journal of Portfolio Management* 20, pp. 68–75.

Wilcox, J., 2000, *Investing by the Numbers* (New Hope: Fabozzi Associates).

13

Removing Long-Only Constraints: 120/20 Investing

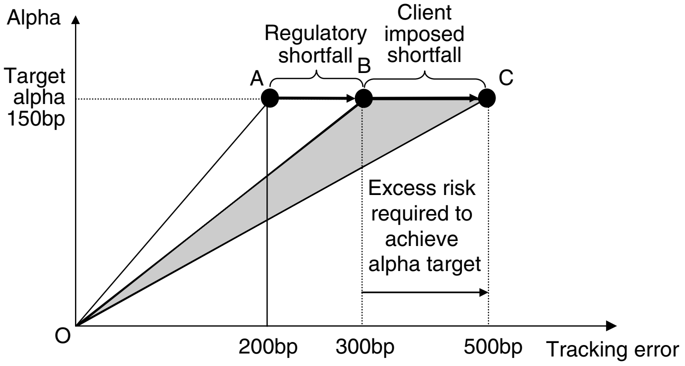
13.1 INVESTMENT CONSTRAINTS

Cynics have long seen constraints as a lawyer's approach to risk management. On the positive side they promote a dialogue between manager and client on risk management and alpha generation, possibly align manager actions with manager skill and protect against a worst case breakdown in risk controls. Constraints are often the direct consequence of the difficult principal agency relationship between manager and client. While the client can directly observe neither skill nor effort, the manager has a strong incentive to pose as skilful (even if they are not) and dial up risks if performance turns against him. While the impossibility to write the perfect contract (in which the agent acts solely in the principal's interest) in a principal agency relationship is well known, investment constraints are partially used to supplement incentive contracts. From the author's perspective, however, many of the constraints used in practice are not in the principal's (plan sponsor's) best interest: rather than reducing the possibility of excess risk taking they increase it. Effectively, most investment constraints (position, turnover, number of stocks, etc) impose a risk management approach of the 1950s as they lack the insights of portfolio theory:

- Nominal constraints (on minimum/maximum weights, maximum duration, etc) are unable to effectively control risk, as they do not take changing volatilities and/or correlations into account.

- No static set of constraints is able to account for the infinite possibilities in portfolio construction. Only a risk measure can.

Figure 13.1 Information ratio shrinkage under alternative constraint sets



Each additional constraint forces the asset manager to take excess risk in order to maintain a given target alpha.

At the same time they reduce breadth in active decisions as they impair the diversification of active bets and sometimes force managers to take risks in areas where the manager does not exhibit skill. As such they are inconsistent with intended risk/return targets (information ratios).

Suppose a given investment process could deliver 150bp of alpha for 200bp in active risk (line OA in Figure 13.1) if left entirely unconstrained. This ends in an information ratio of 0.75. Assume this is the case for hedge funds. At the same time regulatory constraints (most prominently the long-only constraint) bring down the information ratio to 0.5 (150bp of return for 300bp in risk) as line OB shows in Figure 13.1. This can be described as regulatory shortfall. If this shortfall is big enough, it pays for investment managers to move to less regulated investment structures, ie, hedge funds. A whole industry seems to live from the regulatory shortfall. If a given client imposes additional constraints on top of regulatory constraints the risk return ratio is deteriorating further to OC leaving the information ratio at 0.3. For a given alpha target of 150bp the client ends up with much more volatile excess returns than necessary. What was thought to be a safeguard against excess risks turns out to be a road to extra volatility. As such it is no wonder that there is a stronger focus to hedge fund investing and so-called portable alpha solutions. However, it should be clear that not everybody can win from loosening investment constraints. The total amount of alpha is still zero (minus transaction costs) and unconstrained investing can be seen as alpha transfer from constrained to unconstrained investors, but not as a clever way to make everybody better off.

13.2 MEASURING THE IMPACT OF CONSTRAINTS: THE TRANSFER COEFFICIENT

Suppose we are given a universe of n assets where $\boldsymbol{\alpha}$ is an $n \times 1$ vector of asset alphas and $\boldsymbol{\Omega}$ is an $n \times n$ covariance matrix of asset returns. We further assume a standard mean variance objective function to describe how an investor trades off expected return versus risk. For our unconstrained investor the problem is to maximise $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\alpha} - \frac{1}{2}\lambda\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{w}$, where λ denotes the investor's risk aversion and $\boldsymbol{w}$ represents an $n\times1$ vector of portfolio weights. The optimal solution is known to be given by $\boldsymbol{w}_\lambda = \lambda^{-1}\boldsymbol{\Omega}^{-1}\boldsymbol{\alpha}$. The ratio of optimal portfolio weights for different risk aversions is given by $(\boldsymbol{w}_{\lambda_1}/\boldsymbol{w}_{\lambda_2}) = (\lambda_2/\lambda_1)$, ie, solutions across investors of different risk aversion only differ in leverage. All portfolios are perfectly correlated and the efficient frontier is a straight line. In other words, while the relative weights remain the same, individual weights are multiplied by the inverse ratio of risk aversions. For an investor with a risk aversion of 0.01 the optimal weights will be three times as large as for an investor with a risk aversion of 0.03.[1] Information ratio and value added (synonym for utility) of an unconstrained portfolio are given by $\mathrm{IR} = \sqrt{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\alpha}}$ and $U = (1/\lambda)\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\alpha}$.

So far we have assumed that portfolios have been constructed in an unconstrained fashion. The question we want to ask now is, How can we calculate the impact of constraints on portfolio construction? In other words, Can we transfer a given set of alphas to a situation with investment constraints? How much information (ratio) is lost? What does this loss depend on?

Assume we need to constrain the optimal portfolio to be neutral to any given characteristic $\boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is an $n \times 1$ vector of asset characteristics. For example if we require beta neutrality we can simply write $\boldsymbol{\xi} = (\beta_1, \beta_2, \dots, \beta_n)^{\mathrm{T}}$, while for cash neutrality it becomes $\boldsymbol{\xi} = (1, 1, \dots, 1)^{\mathrm{T}}$. The portfolio optimisation problem becomes now

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\alpha} - \tfrac{1}{2}\lambda\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{w} - \upsilon(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\xi}) \tag{13.1}$$

Taking first derivatives with respect to weights and the Lagrange multiplier allows us to solve for optimal portfolio weights[2] as well as the Lagrange multiplier

$$w_\lambda^{\mathrm{c}} = \lambda^{-1}\boldsymbol{\Omega}^{-1}\left(\boldsymbol{\alpha} - \frac{\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}}{\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\xi}^{-1}\boldsymbol{\xi}}\boldsymbol{\alpha}\right) \tag{13.2}$$

$$\upsilon = \frac{\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\alpha}}{\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\xi}} \tag{13.3}$$

Note that as a by-product Equation 13.2 allows us to identify the optimal relative value portfolio (beta neutrality) as well as the optimal timing (active beta positions) given by $w_\lambda^{\text{timing}} = (\upsilon/\lambda)\boldsymbol{\Omega}^{-1}\boldsymbol{\xi}$. In fact comparing Equation 13.2 with the solution for an unconstrained investor reveals that we can achieve the same (constrained) solution in an unconstrained optimisation where we use rescaled alphas $\boldsymbol{\alpha}^{\mathrm{c}} = \boldsymbol{\alpha} - ((\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Omega}^{-1})/(\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\xi}))\boldsymbol{\alpha}$ instead. The information ratio for a constrained investor now becomes

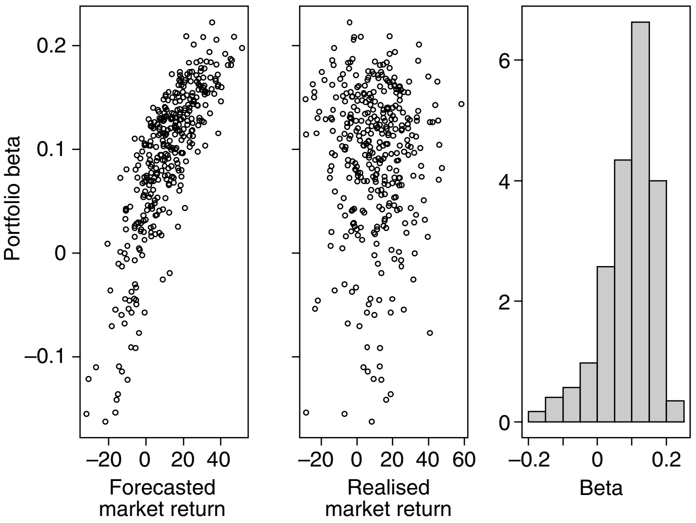
$$\begin{aligned}
\mathrm{IR}^{\mathrm{c}} &= \sqrt{\boldsymbol{\alpha}^{\mathrm{c}\,\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\alpha}^{\mathrm{c}}} \\
&= \sqrt{\left(\boldsymbol{\alpha} - \frac{\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}}{\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\xi}}\boldsymbol{\alpha}\right)^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\left(\boldsymbol{\alpha} - \frac{\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}}{\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\xi}}\boldsymbol{\alpha}\right)} \\
&= \sqrt{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\alpha} - \frac{\bar{\alpha}^2}{\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\xi}}}
\end{aligned} \tag{13.4}$$

Clarke *et al* (2002) introduced the concept of a so-called transfer coefficient, τ. By how much does the information drop if we introduce constraints. In our example the transfer coefficient can be calculated in closed form as[3]

$$\tau = \frac{\mathrm{IR}^{\mathrm{c}}}{\mathrm{IR}} = \sqrt{1 - \frac{\bar{\alpha}^2}{\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\xi}}\frac{1}{\mathrm{IR}^2}} \leqslant 1 \tag{13.5}$$

We can see that an unconstrained investor will always do at least as good as a constrained investor, because $(\bar{\alpha}^2/\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\xi}) \geqslant 0$. Note that $\bar{\alpha}$ equals the alpha for the unrefined alpha forecasts in a portfolio that is neutral to $\boldsymbol{\xi}$ and $(1/\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\xi})$ equals the volatility of a characteristic portfolio with respect to $\boldsymbol{\xi}$. In other words, unless the raw alphas are by accident $\boldsymbol{\xi}$-neutral (in which case there will be no difference between constrained and unconstrained investing) constrained investing will always do worse. Also note that the information ratio remains constant along the alpha efficient frontiers (leverage will not change neutrality to $\boldsymbol{\xi}$). In fact both frontiers are straight lines with different slopes. So the information ratio difference will not depend

Figure 13.2 Simulated market return and beta exposure



on the level of risk aversion. The same is not true for the difference in value added though

$$U - U^{\mathrm{c}} = \frac{1}{\lambda^2}(\mathrm{IR}^2 - \mathrm{IR^c}^2) = \frac{1}{\lambda^2}\left(\frac{\bar{\alpha}^2}{\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\xi}}\right) \tag{13.6}$$

We have seen that mathematically unconstrained investing will in sample (ie, with perfect knowledge of all inputs), always be at least as good as constrained investing. This is mathematically trivial. Unconstrained optimisation will *ex ante* always lead to a larger objective value than constrained optimisation. In practice the concept of is used to measure the impact of arbitrary constraint on *ex post* information ratios. While *ex ante* all constraints are value detracting (as *ex ante* everybody claims to have skill) we will show in the next section, that imposing the right constraints will add value.

13.3 SOME CONSTRAINTS MAKE SENSE

Let us continue with the above example of beta neutrality. What happens if a manager has no market timing skill, but her stock selection signals contain a common market element. We set up a simple simulation exercise.

- Simulate two managers that observe signals on individual stocks.

- Signals contain a common market element.
- Assume a universe of 50 assets with betas varying from 0.5 to 1.5 (to allow indirect market timing via the selection of high beta stocks).
- The information coefficient is 0.1 for each stock. This leaves an unconstraint information ratio of $0.1 \cdot \sqrt{50} = 0.707$.
- Long-only constraints.

We compare two portfolio construction processes. The first process (manager 1) includes a beta neutrality constraint (sum of individual betas multiplied by active weights equals zero), while the second process (manager 2) does not. Let us see what happens to the asset management process for manager 1 in Figure 13.2. Each time the implicit forecast of the market return (hidden in stock selection signals) is positive, the (excess) portfolio beta is positive and vice versa as can be seen from the left panel. However, as the manager has no market timing skill there is no relationship realised market return and excess beta (middle panel). As the market risk premium is positive on average (in about 60% of all cases), the average beta is also positive (right panel). What does this mean for investment performance? We see from Table 13.1 that the unconstrained strategy "learns" that the market has on average a positive risk premium. Excess beta is 0.11 with a *t*-ratio of 6.7, ie, highly significant. A naively calculated information ratio (portfolio return minus benchmark return divided by tracking error) would wrongly find superior skill for the unconstrained (0.66) relative to the beta neutral manager (0.44). However, this is not a reward for skill, but rather excess risk taking (structural risk). A properly calculated (market exposure adjusted) information ratio, however, reverses the result and favours beta neutrality. If left unconstrained, beta risks crowd out alpha generation abilities. Part of the tracking error budget is spent on noisy activities, while it could have been used for alpha generation. Here we have an example where constraints serve a dual purpose. They align manager actions with manager skill and avoid beta repackaging as an investment strategy. After all it is quite common to take structural risks (beta, value, credit, etc) and sell it as alpha generation. Where does this leave us with our discussion about investment constraints? What are the right constraints? Constraints make sense if they do not bind as long as the investment manager does what

Table 13.1 Simulated performance for unconstrained and beta neutral investor

Strategy	IR	β	$\text{IR}_{\text{adjusted}}$
Beta neutral	0.44	0.01 (0.9)	0.44
Unconstrained	0.66	0.11 (6.7)	0.37

was agreed upon. Risk factor neutrality constraints are likely to be a candidate for valuable constraints. The likelihood that stock selectors can add alpha through risk factor timing is low, while at the same time the incentive to repackage beta is high.

13.4 A MORE MATHEMATICAL TREATMENT OF CONSTRAINTS

So far we looked at impact of constraints on total performance. As such the transfer coefficient provided a convenient top line measure. In this section we will specify the costs of constrained investing (especially the long-only constraint) down to the individual security holding level. If we can achieve this we can relate the costs of constraints to individual security characteristics like benchmark weighting or stock market beta. Let us restate the active investment problem from the second section as

$$\max \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\alpha} - \tfrac{1}{2}\lambda \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{w} \tag{13.7}$$

Now we also allow a set of linear equality constraints, that can be summarised under

$$\boldsymbol{A}\boldsymbol{w} = 0 \tag{13.8}$$

where $\boldsymbol{A}$ stands for a $k \times n$ matrix of stock characteristics. Note that k represents the number of equality constraints. If we introduce a cash neutrality, size neutrality and beta neutrality constraint we get $k = 3$ and the exact form is given by

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ s_1 & s_2 & \cdots & s_n \\ \beta_1 & \beta_2 & \cdots & \beta_n \end{bmatrix}, \qquad \boldsymbol{A}\boldsymbol{w} = \begin{bmatrix} \sum_i w_i \\ \sum_i w_i s_i \\ \sum_i w_i \beta_i \end{bmatrix}$$

Note that β_i and s_i are the exposures to a given characteristic, ie, the stock size or benchmark beta. The final constraint is a long-only constraint in disguise as we require

$$\boldsymbol{w} \geqslant -\boldsymbol{w}_{\mathrm{b}} \tag{13.9}$$

and $\boldsymbol{w}_{\mathrm{b}}$ is an $n \times 1$ vector containing the individual benchmark weightings. We can summarise all this in a Lagrange function

$$L = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\alpha} - \tfrac{1}{2}\lambda \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{w} + \boldsymbol{\gamma}^{\mathrm{T}}(\boldsymbol{w} + \boldsymbol{w}_{\mathrm{b}} - \boldsymbol{s}^2) + \boldsymbol{\theta}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{w}) \tag{13.10}$$

where $\boldsymbol{s}^2$ is used for a vector of slack variables. Note that $\boldsymbol{\gamma}$ represents an $n \times 1$ vector of Lagrange multipliers on the individual no short constraint. These multipliers will always be positive as a negative value would actually reverse the constraint. The vector $\boldsymbol{\theta}$ is $k \times 1$ and represents the Lagrange multipliers on our equality constraints. In contrast to the Lagrange multipliers on the long-only constraint, their signs are undetermined. The first-order condition of the following equation tells us that we can write the optimal constrained solution in closed form as a function of the Lagrange multipliers

$$\boldsymbol{w}^{\mathrm{c}} = \lambda^{-1}\boldsymbol{\Omega}^{-1}(\boldsymbol{\alpha} + \boldsymbol{\gamma} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\theta}) \tag{13.11}$$

This is important as these multipliers are a natural way to describe the costs associated with a given constraint. For any binding constraint they describe the shadow price of relaxing these constraints in terms of units of the objective function won. In fact Equation 13.11 is only one among many KKT conditions for solving a quadratic programme.[4] At this stage we do not intend to ask ourselves how to optimally determine the Lagrange multipliers. However, virtually all optimisation packages routinely provide the Lagrange multipliers of a constrained optimisation problem. There is no need for the reader to replicate the calculations done by sophisticated packages like CPLEX or NuOPT. All they need to do is to store, use and interpret the Lagrange multipliers in the way we suggest in this chapter. Note that we can use this framework to express liquidity costs. The inability to enter a position works like a constraint and as such we can relate the lack of liquidity to its costs.

We start with the well-known expression for implied returns. For a given vector of constrained portfolio weights, $\boldsymbol{w}^{\mathrm{c}}$ the implied alphas $\boldsymbol{\alpha}^{\mathrm{c}}$ can be defined as $\lambda\boldsymbol{\Omega}\boldsymbol{w}^{\mathrm{c}}$. In other words, the required alpha for a given asset equals its marginal contribution to risk times risk aversion. Multiplying both sides of Equation 13.12 with $\lambda\boldsymbol{\Omega}$ and using the above identity we can rewrite the first-order condition to arrive at[5]

$$\boldsymbol{\alpha}^{\mathrm{c}} = \boldsymbol{\alpha} + \boldsymbol{\gamma} + A^{\mathrm{T}}\boldsymbol{\theta} \tag{13.12}$$

where $\boldsymbol{\alpha}^{\mathrm{c}}$ denotes the vector of implied returns of the constrained solution. Economically this means that we now have a way to decompose the alphas that have been distorted through the introduction of constraints into the origins of these distortions. For the ith alpha Equation 13.12 can be expressed more transparently as

$$\alpha_i^{\mathrm{c}} = \alpha_i + \gamma_i + \theta_{\mathrm{c}} \cdot 1 + s_i\theta_{\mathrm{s}} + \theta_\beta \beta_i \tag{13.13}$$

where θ_{c}, θ_{s} and θ_β reflect the multipliers on cash, size and beta neutrality constraints. It is worthwhile spending some time on Equation 13.13 as it essentially provides valuable insight how raw alphas (α_i) are being transformed during a process of constrained optimisation into implied alphas (α_i^{c}). The first adjustment to raw alphas is made for hitting the lower barrier on individual position limits. If a lower limit (eg, −3.5% benchmark weight) for the ith stock is hit, the Lagrange multiplier will be positive $\gamma_i \geqslant 0$ and as such the implied alpha (α_i^{c}) of this stock is larger than the original (α_i). This correction is needed to make a larger short position unattractive. Constraints implicitly change the forecast our valuation models make and therefore create inefficiencies. The remaining equality constraints are broken down to the individual security level by using the security exposure to a given constraint. How does this work? Suppose we have an exceptionally high alpha to a large cap stock, but the size constraint forces us to remain size neutral. In this case we are most likely forced to underweight another large cap stock (even though it might have decent positive returns) as the overweight needs to be funded by underweight. In short: while a given alpha has the same contribution on active returns irrespective of size, it will have a much larger (distorting effect) on constraints for a large cap stock.

13.5 CONSTRAINTS, ALPHA AND VALUE ADDED

The previous section essentially reviewed the work by Grinold and Eaton (1998). Following their approach we used the relationship expressed in Equation 13.13 to explain the distorted alphas as a function of constraints. This allows us to answer questions like: what harm can constraints do to your forecasts? The more interesting question, however, is to ask: what harm can constraints do to value added, ie, investors' utility. Remember that constraints might create distortions in implied alphas, but could still have little impact on value added. This section will explain our approach. Let us focus on the alpha differential first. We suggest using the first order KKT conditions in Equation 13.12 to directly break down the difference in portfolio alpha for the constrained and unconstrained portfolios. In other words we want to decompose the difference between the returns of an unconstrained portfolio (α) and a constrained portfolio (α^{c}) into a function of Lagrange multipliers. We start with the definition of the alpha differential

$$\alpha - \alpha^{\mathrm{c}} = (\boldsymbol{w} - \boldsymbol{w}^{\mathrm{c}})^{\mathrm{T}}\boldsymbol{\alpha} \tag{13.14}$$

Now we substitute Equation 13.12 into Equation 13.14 and use $\boldsymbol{w} = \lambda^{-1}\boldsymbol{\Omega}^{-1}\boldsymbol{\alpha}$ to arrive at

$$\begin{aligned}\alpha - \alpha^{\mathrm{c}} &= (\lambda^{-1}\boldsymbol{\Omega}^{-1}\boldsymbol{\alpha} - \lambda^{-1}\boldsymbol{\Omega}^{-1}(\boldsymbol{\alpha} + \boldsymbol{\gamma} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\theta}))^{\mathrm{T}}\boldsymbol{\alpha}\\ &= (\lambda^{-1}\boldsymbol{\Omega}^{-1}(-\boldsymbol{\gamma} - \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\theta}))^{\mathrm{T}}\boldsymbol{\alpha}\\ &= (-\boldsymbol{\gamma}^{\mathrm{T}} - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{A})\lambda^{-1}\boldsymbol{\Omega}^{-1}\boldsymbol{\alpha}\\ &= -(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{w} + \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{w})\end{aligned}$$

In other words, the (positive) difference between unconstrained and constrained returns is a function of the Lagrange multipliers and as such can be decomposed into its origins.

Expanding the above expression we get the following breakdown

$$\alpha - \alpha^{\mathrm{c}} = -\sum_i w_i\gamma_i - \theta_c\sum_i w_i - \theta_{\mathrm{s}}\sum_i w_i s_i - \theta_\beta\sum_i w_i\beta_i \tag{13.15}$$

For example, the first term, $-\sum_i w_i\gamma_i$ represents the contribution of the long-only constrained. It is clearly positive as γ_i is positive only for those stocks where the long-only constraint is binding, ie, where w_i is negative.

The ultimate objective of any analysis regarding the costs of constraints is to quantify the impact constraints have on the loss in utility relative to an unconstrained solution. This has not been addressed in the existing literature. Using the same framework as above we will now show how to decompose the difference in value added into its contributors. Let us first substitute $\boldsymbol{\alpha} = \lambda\boldsymbol{\Omega}\boldsymbol{w}$ into the difference in value added given in Equation 13.16

$$U - U^{\mathrm{c}} = (\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\alpha} - \tfrac{1}{2}\lambda\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{w}) - (\boldsymbol{w}^{\mathrm{cT}}\boldsymbol{\alpha} - \tfrac{1}{2}\lambda\boldsymbol{w}^{\mathrm{cT}}\boldsymbol{\Omega}\boldsymbol{w}^{\mathrm{c}}) \tag{13.16}$$

where unconstrained and constrained utility are denoted by U and U^{c} respectively. The difference in utility is then more compactly written as

$$\begin{aligned} U - U^{\mathrm{c}} &= \tfrac{1}{2}\lambda(\boldsymbol{w} - \boldsymbol{w}^{\mathrm{c}})^{\mathrm{T}}\boldsymbol{\Omega}(\boldsymbol{w} - \boldsymbol{w}^{\mathrm{c}}) \\ &= \tfrac{1}{2}\lambda(\lambda^{-1}\boldsymbol{\Omega}^{-1}(\boldsymbol{\gamma} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\theta}))^{\mathrm{T}}\boldsymbol{\Omega}(\lambda^{-1}\boldsymbol{\Omega}^{-1}(\boldsymbol{\gamma} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\theta})) \\ &= \frac{1}{2\lambda}(\boldsymbol{\gamma} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{\Omega}^{-1}(\boldsymbol{\gamma} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\theta}) \end{aligned} \tag{13.17}$$

Second, we can view this as risk term $\psi^2 \equiv 2\lambda(U - U^{\mathrm{c}})$.[6] The reader can recognise it as linearly homogeneous and hence decompose it into its marginal contributions, ie

$$\begin{aligned} \psi &= \boldsymbol{\gamma}^{\mathrm{T}}\frac{\mathrm{d}\psi}{\mathrm{d}\boldsymbol{\gamma}} + \boldsymbol{\theta}^{\mathrm{T}}\frac{\mathrm{d}\psi}{\mathrm{d}\boldsymbol{\theta}} \\ &= \boldsymbol{\gamma}^{\mathrm{T}} \cdot \frac{\boldsymbol{\Omega}^{-1}(\boldsymbol{\gamma} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\theta})}{\psi} + \boldsymbol{\theta}^{\mathrm{T}}\frac{\boldsymbol{A}\boldsymbol{\Omega}^{-1}(\boldsymbol{\gamma} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\theta})}{\psi} \end{aligned} \tag{13.18}$$

Finally, multiplying both sides by ψ again we arrive at

$$U - U^{\mathrm{c}} = \frac{1}{2\lambda}(\underbrace{\boldsymbol{\gamma}^{\mathrm{T}} \cdot \boldsymbol{\Omega}^{-1}(\boldsymbol{\gamma} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\theta})}_{\substack{\text{Contribution of} \\ \text{no short constraint}}} + \underbrace{\boldsymbol{\theta}^{\mathrm{T}} \cdot \boldsymbol{A}\boldsymbol{\Omega}^{-1}(\boldsymbol{\gamma} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\theta})}_{\substack{\text{Contribution of} \\ \text{equality constraint}}}) \tag{13.19}$$

This gives us a compact expression for the difference in utility as a function of the respective Lagrange multipliers.[7]

13.6 THE MECHANICS OF 120/20 INVESTING

To analyse the benefit of partially relaxing the long-only constraint, we first lay out the long–short optimisation formulation used in our example. For formulation convenience, we denote securities' active holding positions as

$$w_i = w_{\mathrm{l},i} - w_{\mathrm{s},i} - w_{\mathrm{b},i}, \quad w_{\mathrm{l},i} \geqslant 0,\ w_{\mathrm{s},i} \geqslant 0 \tag{13.20}$$

where $w_{\mathrm{l},i}$ represents a security's long position while $w_{\mathrm{s},i}$ denotes its short position, again $w_{\mathrm{b},i}$ is the benchmark weight. Since the partially relaxed long-only portfolio, eg, 120/20, allows short position as long as the portfolio's total leverage exposure is within a prespecified range, we need to replace the long-only constraint by a leverage constraint which specifies the portfolio's maximum leverage exposure L

$$\sum_i w_{\mathrm{l},i} + \sum_i w_{\mathrm{s},i} \leqslant 1 + L \tag{13.21}$$

For a long side of 120 we need to short (borrow against a lending fee and sell) stock worth 100. We also might want to introduce integer variables to prevent simultaneous long and short for same securities. We use the following constraints to achieve this

$$b_i m \geqslant w_{1,i}, \qquad (1 - b_i)m \geqslant w_{\mathrm{s},i} \tag{13.22}$$

where m is a "large" number. Note that Equation 13.24 operates like a switch. If $b_i = 1$, then the short position of security i will be forced to equal zero, otherwise, the long position will equal zero. The size neutrality and beta neutrality constraints are the same as before. Although the above formulation allows us to find the optimal active holding position of a relaxed problem, it also makes the explanation of the shadow costs unclear because of the separation of the long and short position. Another downside is that the Lagrange problem becomes non-differentiable exactly because of the existence of the integer variable.

To make our utility decomposition methodology work for the partially relaxed long-only problem, we propose a two-step optimisation approach to calculate the shadow costs of constraints. In the first step, we solve the optimisation problem with the help of the binary variables to achieve a portfolio satisfying the leverage and other equality constraints. We then prespecify the sign of the holding positions of all the securities according to the optimal value of the binary variables and resolve an equivalent problem. The equality constraints can be summarised as

$$\left.\begin{aligned}
&\begin{bmatrix} l_1 & l_2 & \cdots & l_n \\ 1 & 1 & \cdots & 1 \\ s_1 & s_2 & \cdots & s_n \\ \beta_1 & \beta_2 & \cdots & \beta_n \end{bmatrix}, \quad Aw = \begin{bmatrix} \sum\limits_i S_i w_i l_i \\ \sum\limits_i S_i w_i \\ \sum\limits_i w_i S_i \\ \sum\limits_i w_i \beta_i \end{bmatrix} \\
&B = \begin{bmatrix} 1 + L - \sum\limits_i wb_{,i}\, l_i \\ 0 \\ 0 \\ 0 \end{bmatrix}
\end{aligned}\right\} \tag{13.23}$$

The first constraint is the leverage constraint which requires the portfolio's leverage to equal L. The coefficients of the leverage constraint are defined as

$$l_i = \begin{cases} 1 & \text{if } b_i = 1 \\ -1 & \text{if } b_i = 0 \end{cases} \tag{13.24}$$

where b_i is the value of the integer solution of the model with integer decision variables. In addition to the above equality constraints, the inequality constraints specify the sign of the active positions based on the value of b_i

$$w_i + w_{\mathrm{b},i} \geqslant 0 \quad \text{if } b_i = 1$$
$$w_i + w_{\mathrm{b},i} \leqslant 0 \quad \text{if } b_i = 0$$

We can now find the dual variable for each constraint and analyse constraints' effects on alpha distortion and utility loss. In fact, this two-step optimisation approach assumes that the optimiser uses the branch and bound approach to tackle the integer problem, and the second step simply takes the optimal integer solutions to find the meaningful shadow costs for each constraint.

13.7 CONCLUSION

This chapter provided a closer look at the costs of constraints and suggested a new methodology to measure the impact of individual constraints on an investor's value added. Rather than calculating the decay in information ratio, we showed how to use Lagrange multipliers in order to decompose the difference in value added into the contributions coming from alternative constraints. This differs considerably from current methods that either focus on a headline measure of information ratio shrinkage (also called transfer coefficient) or express the impact of constraints as distortions in implied alphas. While the later approach can answer questions like – "What harm can constraints do to your forecasts?" – the more interesting question to us is to ask: "What harm can constraints do to value added, ie, investor's utility?" Constraints might create distortions in implied alphas, but could still have little impact on value added, as we have seen from the size constraint in our setting. We also suggest using the decomposition of value added into its constraint related parts as a diagnostic measure. As real alphas show no correlation to firm specific characteristics (otherwise they would be called beta) constraints on various betas will show little impact on the optimality of the solution. Size, beta or value neutrality constraints only become disruptive when an investment process becomes dominated by taking factor bets. In this case it is not clear that one should take off the constraints. In fact this has been the reason why constraints were there in the first place, ie, align portfolio construction with skill. Finally we would like to mention that this kind of analysis can only be performed for an asset management process that provides explicit signals (rather than directly choosing weights) and uses portfolio optimisation. In other words, quantitative managers will benefit most.

[1] Note that $\lambda = (\alpha/\sigma^2)$. For $\lambda = 0.03$ this means, that an investor targeting an alpha of 1 is willing to take a risk (standard deviation) of 5.75.

[2] To see if $\boldsymbol{W}_\lambda^{\mathrm{cT}}\xi = 0$, we calculate

$$\left(\boldsymbol{\alpha}^{\mathrm{T}} - \boldsymbol{\alpha}^{\mathrm{T}}\frac{\boldsymbol{\Omega}^{-1}\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}}{\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\xi}}\right)\lambda^{-1}\boldsymbol{\Omega}^{-1}\boldsymbol{\xi} = \lambda^{-1}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\xi} - \lambda^{-1}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\xi}\frac{\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\xi}}{\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\xi}} = 0$$

[3] Another way to express the transfer coefficient in an *ex ante* context is to calculate the correlation between the vector of constrained and unconstrained weights, given as $\boldsymbol{\tau} = (\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}^{\mathrm{c}} / \operatorname{norm}(\boldsymbol{w}) \operatorname{norm}(\boldsymbol{w}^{\mathrm{c}}))$.

[4] See Fletcher (1987) for an excellent resource on how to practically solve optimisation problems.

[5] See Grinold and Eaton (1998).

[6] See Grinold (2006, Equations 12.5–12.9) for an independent derivation within a similar focus.

[7] The reader can check that by expanding Equation 13.21 we get Equation 13.18 again.

REFERENCES

Clarke, R., H. de Silva and S. Thorley, 2002, "Portfolio Constraints and the Fundamental Law of Active Management", *Financial Analysts Journal* 58(5), pp. 48–66.

Fletcher, R., 1987, *Practical Methods of Optimization* (Chichester: John Wiley & Sons).

Grinold, R. C., 2006, "Implementation Efficiency", *Financial Analysts Journal* 61(5), pp. 52–63.

Grinold, R. C., and K. A. Eaton, 1998, "Attribution of Performance and Holdings", in W. T. Ziemba and J. M. Mulvey (eds), *Worldwide Asset Liability Modeling* (Cambridge University Press).

14

Performance-Based Fees, Incentives and Dynamic Tracking Error Choice

Mutual funds and corporate pension funds are increasingly using incentive fees, or performance-based fees (PBFs), to reward their fund managers. Janus Capital Group Inc established performance incentives for the managers of 13 of its 59 funds in September 2005. Vanguard and Fidelity are just another two examples of investment companies which use incentive fees. Even in the absence of explicit performance-based fees, Brown *et al* (1996) and Chevalier and Ellison (1997) have shown that an implicit performance-based compensation structure arises from proportional fees as a result of the fact that net investment flows into funds respond strongly to recent performance.

The effects of performance-based fees on investment decisions have been documented in a number of papers. Grinblatt and Titman (1989) apply option pricing theory to analyse the manager's risk incentives in a single-stage framework. They find that improperly designed PBF contracts create incentives for gaming by varying the risk of the fund. Carpenter (2000) examines the optimal dynamic investment policy for a risk-averse fund manager and finds that the convexity of the option-like compensation structure can lead the manager to dramatically increase volatility in some circumstances. Goetzmann *et al* (2003) focus primarily on valuing claims on a hedge fund's assets and show that convexity gives rise to a risk-shifting incentive on the part of the money manager. Ferguson and Leistikow (2003) also use a multi-period framework, but they treat it as a sequence of myopic single-period problems. Basak *et al* (2007) examine mutual fund managers' risk-taking behaviour when compensation is a function of fund value, which in turn depends on flows.

Although these papers contain many useful insights, they focus either on a single-period model or assume a restrictive continuous-time framework. In either case, the model is motivated more by analytic tractability than by realistic investment considerations. For example, many investors expect their active risk relative to the benchmark to be maintained within a range. Investors also have limited tolerance for underperformance and are likely to redeem their investments from funds that trail their benchmarks significantly. Ippolito (1992) finds that poor fund performance results in large money outflow, while good performance leads to significant inflow of new money. Khorana (1996) shows that up to two years of below average performance significantly increases the probability that the mutual fund manager will be fired. We shall address these considerations in a multi-stage framework using stochastic programming techniques.

Hodder and Jackwerth (2007) have analysed a fund manager's risk-taking behaviour when faced with a liquidation barrier. However, their work mainly focuses on asymmetric compensation structure. The focus of this chapter is the behaviour of active portfolio managers under different performance-based compensation structures, the most common examples of which are proportional fees and incentive fees, or a combination of these.[1] In practice, the incentive fee can be of either the asymmetric type (no loss participation) or fulcrum type (loss participation). Recognising that the incentive fee is an option contingent on the performance of the fund, the problem of determining a fund manager's optimal active risk bears some relation to the problem of valuing a derivative.[2] However, the investment manager cannot be allowed to hold a short position in this performance option, synthetic or otherwise, as this would subvert the investor's intended purpose in paying a performance fee. The portfolio manager cannot be allowed to delta hedge the option in their private portfolio.

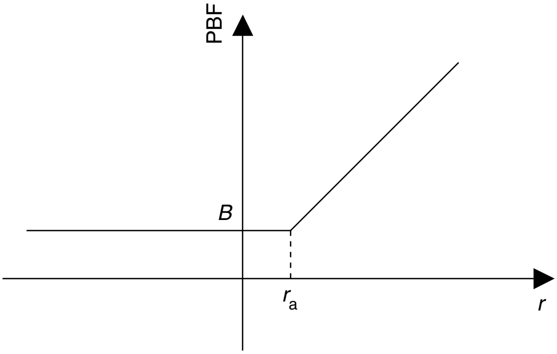
Many papers have considered the optimal contract design problem in a multi-stage game-theoretic framework. Das and Sundaram (1999) and Stremme (2001) have attacked the problem from a signalling perspective. Their work centres around what they call "fee speech", ie, what does a given fee structure tell us about the manager? Is there a pooling equilibrium in which every manager (good as well as bad) offers the same contract or is the market outcome separating, ie, can we tell whether a manager has skill by looking at their fee structure?

Traditional analysis of performance-based fees does not consider the dynamic nature of active portfolio management. In reality, fund managers are more concerned about the renewal of their lucrative long-term contracts than about the marginal benefits they may gain by varying risk in the short term. This chapter addresses the effect of investor intolerance of negative fund performance on the valuation of incentive fees and the manager's optimal risk policy response in this setting. In our model, if the fund severely underperforms the benchmark, then the manager is fired. Specifically, we assume that the investor has a minimum performance tolerance barrier, which they enforce based on periodic performance reports. At the end of any reporting period, the management contract will be terminated if performance is below the investor tolerance barrier. We refer to this as a knockout, or down-and-out, performance barrier. The aim of this chapter is to address the multi-stage portfolio management problem for a utility-maximising investment manager in the presence of a discrete knockout barrier or down-and-out performance barrier. Even though they are determined exogenously, our underperformance tolerance assumptions are consistent with industry reality, especially in the case of mutual funds, for which market forces play an important role. Since the performance history of mutual funds is public information, investors are well informed and willing to act on news of poor returns.

14.1 REVIEW OF THE SINGLE-STAGE INCENTIVE-FEE MODEL

We start with a review of the single-period model in which the money manager is subject to some form of performance-based compensation structure. While investors and portfolio managers use different performance-based fee contracts, which vary widely in terms of complexity and sophistication, we focus on two popular contract types. Hedge funds typically charge investors a fixed percentage of assets under management plus an asymmetric incentive fee, while mutual funds typically charge a fixed percentage plus a symmetric "fulcrum fee" (at least in the US) as incentive.

Figure 14.1 Portfolio manager's payout under asymmetric fee compensation structure as a function of realised return



B, flat fee; r_a is the hurdle point for the asymmetric fee.

Figure 14.1 gives the most commonly used compensation structure for hedge funds. It contains a proportional fee and an incentive fee. The proportional fee takes the form $B = k_1 W$, where k_1 is typically 1–2% and W denotes assets under management (AUM) at the end of the period. The incentive fee is typically a proportion of the outperformance of the portfolio over a prespecified target (high watermark returns in the case of hedge funds). However, other targets are common, such as a hurdle point r_h above a benchmark.

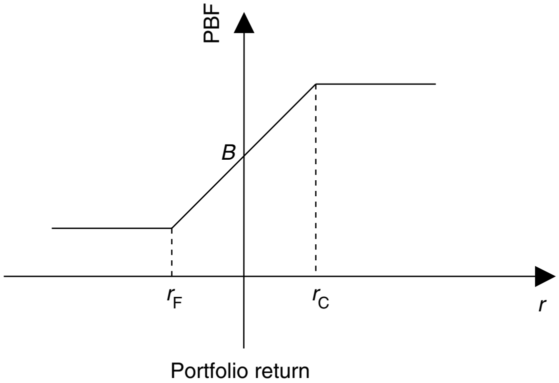
Denote by r_M and r_B the returns of the managed portfolio and the benchmark portfolio respectively. Then the managed fund's relative return[3] with respect to the benchmark at the end of the investment period can be expressed as

$$r_1 = \frac{1 + r_M}{1 + r_B} - 1$$

If the manager outperforms the target return by an amount $r_1 - r_h$, the bonus compensation equals $k_2(r_1 - r_h)W_0$, where $0 < k_2 < 1$ is the fraction of the excess return over the hurdle rate that goes to the manager and W_0 is the initial wealth of the fund. In all other cases, the manager only collects a proportional fee, ie, a fee that is a certain percentage of the fund's assets W at the end of the investment period. The hedge-fund manager's compensation at the end of the period has the form

$$F(r_1) = k_1 W + \max[0, k_2(r_1 - r_h)W_0]$$

Figure 14.2 Portfolio manager's payout as a function of realised return under fulcrum-fee compensation structure



B is the propositional fee, r_F and r_C are the floor and cap rates for the fulcrum-fee contract.

From Figure 14.1 we see that the fund manager's performance-based compensation incorporates a call option. Since this fee schedule effectively limits downside risk for the portfolio manager, we would expect the manager to prefer riskier portfolios to maximise their personal utility. Grinblatt and Titman (1989) have shown that performance-based compensation contracts as in Figure 14.1 provide an incentive for fund managers to deviate from the risk level which is optimal for the investor. Carpenter (2000) also points out that the convexity of the option makes the manager seek payouts that can lead to dramatic increases in portfolio's volatility.

To protect public investor's interest, the incentive-fee structures in the US mutual fund industry are regulated according to the Investment Company Amendments Act of 1970, an amendment to the Investment Company Act of 1940. This amendment requires all mutual fund managers wishing to charge performance-based fees to adopt a fulcrum-fee structure. Figure 14.2 illustrates a typical fulcrum-fee contract, which takes the form of a base proportional fee, B, plus an adjustment for outperforming or underperforming the benchmark. With a fulcrum fee, the compensation is symmetric around a chosen index; so that any increases in fees for performance in excess of the benchmark, r_B, must be matched by decreases in fees for performance that falls short of the benchmark by a like amount.

In the case when the return of the managed fund equals the benchmark return, ie, $r_1 = 0$, the manager's payout will only contain the propositional fee $k_1 W$. The incentive portion of the fulcrum fee has a cap hurdle rate, r_C, and a floor hurdle rate, r_F. To satisfy the legal requirement of symmetry of fulcrum fees, we must have $r_C = -r_F$. If the managed portfolio's realised return is greater than r_C, the "cap" fee will be $k_1 W + k_2 r_C W_0$. Symmetrically, if the managed portfolio underperforms the benchmark by an amount greater than r_F, the manager's "floor" fee level is $k_1 W + k_2 r_F W_0$. When relative return is in the intermediate range $[r_F, r_C]$ the performance fee increases linearly from the floor to the cap. The manager's payout can be expressed as

$$F(r_1) = k_1 W + k_2 W_0 \min[r_C, \max[r_F, r_1]]$$

Although the symmetric structure of the fulcrum fee mitigates the manager's risk-shifting incentive when the fund's return is near the benchmark, there is still a significant incentive to deviate from the client's preferred level of risk when their relative return is near r_F or r_C. In the case when the managed portfolio's performance is far behind the benchmark, the manager has a strong incentive to increase the fund's risk level because their downside risk is limited; they will also play it safe and act more like an index fund if they are considerably ahead of the benchmark. This applies in a single-period framework, because there is no opportunity for the investor to punish the manager for deviating from the optimal contract. We will relax this assumption in the next section.

14.2 MULTI-PERIOD SETTING AND THE PORTFOLIO MANAGER'S OPTIMISATION PROBLEM

In the remainder of this chapter we formulate a multi-stage portfolio management problem with incentive fees and a knockout barrier for poor performance. The knockout feature is consistent with business reality, especially for institutional investors. Heinkel and Stoughton (1994) observe that almost all major institutional investors pay professional portfolio evaluation firms, such as Frank Russell or Wilshire Associates, to monitor and report managers' performance on a regular basis. Since the performance of the money managers (usually measured annually) can easily be observed and the cost of transferring money from one manager to another is not high, significantly underperforming the benchmark will result in redemptions from the fund. In the case of mutual funds, where performance information is publicly available, investors are well informed and will exit a fund if performance is poor compared with its competitors.

To consider an investor's performance tolerance in a dynamic investment environment, our multi-period model assumes there is a minimum performance knockout barrier for the investor, and they can observe the performance of the fund dynamically. At the end of each period, the manager will be fired if the fund's relative performance is below the knockout barrier. A more formal specification of the model is as follows. Initially, the investor turns over a sum of money, W_0, to the fund manager and delegates fund investment decisions to them over a certain length of time. Assume the portfolio will be rebalanced T times over the investment horizon and the investor can observe the fund's performance at the end of each period (labelled as $1, 2, \dots, T$). At the beginning of each period, the manager will make a decision to maximise their expected utility at the end of the investment horizon, subject to any constraint imposed by the investor. The final payout depends on the fund's performance over the investment horizon in a manner specified in the agreed compensation contract.

There is a prespecified benchmark against which the performance of the managed portfolio will be evaluated. Let

$$r_t = \frac{1 + r_{t,\mathrm{M}}}{1 + r_{t,\mathrm{B}}} - 1$$

be the relative return on the investment during period t, where $r_{t,\mathrm{M}}$ and $r_{t,\mathrm{B}}$ denote the returns of the managed and the benchmark portfolios respectively. We assume that the manager will maximise their personal utility by adjusting the portfolio's active risk level, ie, the volatility of the active returns. In our model, the level of active risk σ chosen will determine the entire distribution of return; in particular, the active[4] expected return α will be a function of active risk.

We must explain the relationship between active risk σ and expected active return α before we formalise the manager's utility maximisation problem. A key statistic for measuring the manager's investment skills is the information ratio, defined as $\mathrm{IR} \equiv \alpha/\sigma$. This is a concept similar to the Sharpe ratio. A larger information ratio indicates a higher active return per unit of active risk. The relationship between active risk and active return is specified by active efficient frontier $\alpha(\sigma)$. For a given level of active risk, $\alpha(\sigma)$ is the largest alpha the manager can achieve. In another words, the active efficient frontier defines the quality of the opportunities available to a manager. Each active manager will have their own active efficient frontier. The active efficient frontier of a good manager will dominate that of a poor one. Through out this chapter, we assume that the manager's frontier is static, ie, that the manager's investment skill will not change during the investment horizon.

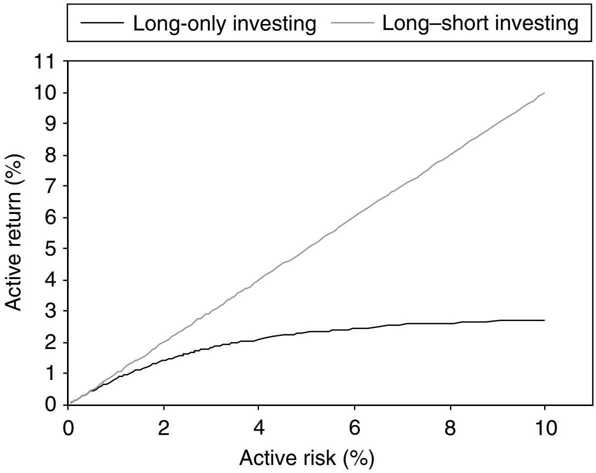
The active efficient frontier for an unconstrained investor is a straight line through the origin. In this case, the information ratio of the optimal portfolio is independent of active risk. This situation is fairly typical for a hedge-fund investor. However, for a constrained investor, for example a typical mutual-fund investor with a long-only constraint (more precisely, one who cannot borrow), the frontier becomes concave and the information ratio decreases as the portfolio's active risk increases. Since a long-only requirement incurs inequality constraints, there is no explicit formula for the loss in efficiency. Grinold and Kahn (2000) carried out a numerical experiment to explore the impact of such constraints. They summarised their conclusions in the form of an approximate formula showing how much IR is typically lost because of long-only constraints as a function of the number of assets and the active risk. The sample efficient frontier we used to illustrate the effects of long-only constraints is based on Grinold and Kahn's simple model, which estimates the active efficient frontier with long-only constraints by

$$\alpha(\sigma) = 100 \cdot \mathrm{IR} \cdot \left\{ \frac{[1 + \sigma/100]^{1-\gamma(N)} - 1}{1 - \gamma(N)} \right\}$$

where IR is the information ratio without long-only constraints (assumed to be 1.0), σ is security's residual risk, N is number of securities and $\gamma(N) = (53 + N)^{0.57}$. This specification is used for asset-based-fee and fulcrum-fee cases (typically with long-only constraint). For hedge funds we use $\alpha = \mathrm{IR} \cdot \sigma$ instead.

In the examples to follow, we take our active efficient frontiers to be as in Figure 14.3. The long-only frontier is based on the Grinold–Kahn model with 500 assets and an unconstrained IR of 1. The long–short frontier is based on an IR of 1: we imagine this to reflect an unconstrained strategy based on the same information set as the

Figure 14.3 Active efficient frontiers for portfolios with and without long-only investment constraints



long-only frontier. The loss of efficiency as risk increases is clearly visible in the long-only case. The long-only frontier represents the mutual fund case, and will always be coupled with a fulcrum-fee structure. The long–short frontier represents the hedge-fund case, and will always be coupled with the asymmetric fee structure.

Our analysis to follow assumes that r_t is normally distributed; specifically, that $r_t \sim N(\alpha_t(\sigma_t), \sigma_t^2)$. The normality assumption is largely one of convenience because the stochastic programming method we use applies equally well to any return distribution. We also assume that r_t is conditionally independent of the past given σ_t.

If the fund's cumulative relative performance from time 0 to time t, $R_t \equiv \prod_{t=1}^{t}(1 + r_t)$, is below a predetermined knockout barrier level,[5] $R_{t,\text{barrier}}$, the portfolio manager will be fired. Let I_t denote the investor's knockout indicator

$$I_t = \begin{cases} 1 & \text{if } R_t \geqslant R_{t,\text{barrier}} \\ 0 & \text{otherwise} \end{cases}$$

$I_t = 1$ indicates that the manager is not fired and may continue to run the fund through the next period, period $t + 1$. If the manager has not been fired at the end of period T, they will be paid according to a prespecified compensation structures and the realised relative return.

Otherwise, the manager receives only the fixed proportional fee of the initial fund wealth, pro-rated according to the number of periods over which they managed the fund before they were fired. We assume that the fee is received upon termination of the management contract and is invested at the risk-free rate until time T.

We assume that the manager's only decision in each period $t \in [1, \dots, T]$ is the choice of the risk level $\sigma_t \in [\sigma_{t,\mathrm{L}}, \sigma_{t,\mathrm{H}}]$, where $[\sigma_{t,\mathrm{L}}, \sigma_{t,\mathrm{H}}]$ is the risk range specified for period t. In our model, σ also implicitly determines the portfolio's expected excess return α. Further, assume the portfolio manager's utility is based on the total fees accrued by time T (including fees received before time T, which are invested at the risk-free rate) and exhibits constant relative risk aversion (CRRA)

$$U(W_T) = \frac{W_T^{1-\gamma}}{1-\gamma}, \quad \gamma > 0$$

Given the above setting, the portfolio manager's utility maximisation optimisation problem[6] can be expressed as

$$\max_{\sigma} E[U(F(R))], \quad \text{where } \sigma_t = \sigma_t(R_t) \text{ for } t = 1, 2, \dots, T$$

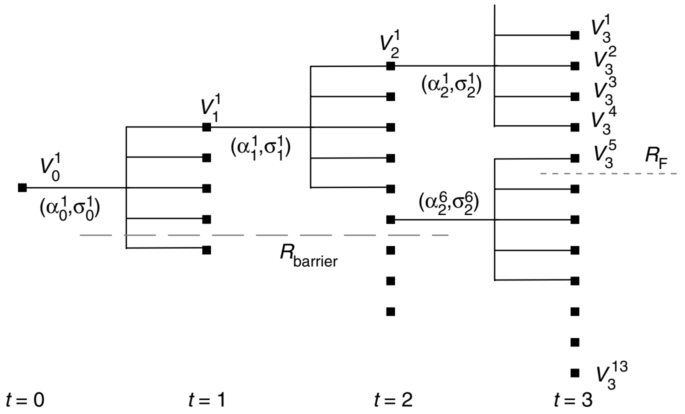
where F is the total accrued fees at the end of the horizon, which depends on the history of the fund's performance, R, and in particular on the investor's knockout indicator process, I. The maximal expected utility is a function of the contract parameters Φ, such as the initial assets under management, W_0, the time-dependent knockout barrier R_{barrier} and the different hurdle rates and fees.

The complexity of many real-world problems often means that we have to abandon hope of an elegant closed-form solution as our mathematical models are refined. Because the model presented here incorporates many of the practical nuances of investor demands, manager motivation and investment skill, it has no closed-form solution. This is the price we pay for the realism of the model. We therefore developed a stochastic programming algorithm to solve the portfolio manager's active risk optimisation problem.

14.3 SCENARIO TREE GENERATION AND THE OPTIMISATION ALGORITHM

We use dynamic programming to solve the portfolio manager's dynamic tracking error optimisation problem. This section illustrates the scenario tree-generation process and a solution algorithm

Figure 14.4 A simple three-stage five-kid scenario tree



The fund's investment horizon is separated into three stages. For each "parent" node (where decisions are taken) there are five kids (where decisions are evaluated and new decisions are formed). R_{barrier} is the knockout cumulative relative return level and R_{f} is the floor level of the asymmetric compensation parameter.

we proposed. Note that the complexity of modelling a real world problem often (mostly) does not allow the elegance of closed-form solutions. While some readers might find this to be a disadvantage, it is commonplace in most financial applications ranging from portfolio optimisation to option pricing. Similar to binomial or trinomial option pricing, we first generate a multinomial scenario tree to represent the possible realisation of the managed portfolio's relative performance at the end of each stage, and then use backward-induction method to find the optimal solution.

The contract horizon is divided into T stages, each corresponding to a managerial evaluation period. For simplicity, we assume the managed portfolio has static active efficient frontier and constant risk range $[\sigma_{\text{L}}, \sigma_{\text{U}}]$ specified by the investor over the contract horizon. The starting point's cumulative relative performance is set to 1. Let P be the probability distribution of the gross relative return, R, over a single investment stage associated with a node in the scenario tree. Denote by R_{L} and R_{U} the lower and upper ε percentile level (eg, 1%) of the gross relative return assuming the maximum tracking error decision is taking and the alpha equal to 0, ie

$$P[R \leqslant R_{\text{L}} \mid \alpha = 0,\ \sigma = \sigma_{\text{U}}] = \varepsilon$$

$$P[R \geqslant R_{\text{U}} \mid \alpha = 0,\ \sigma = \sigma_{\text{U}}] = \varepsilon$$

To represent the investment uncertainty over a single stage, K discrete samples are taken from $[R_{\mathrm{L}}, R_{\mathrm{U}}]$. This is accomplished by multiplying the initial position by a factor u_k

$$u_k = R_{\mathrm{L}}\mathrm{e}^{(k-1)\Delta} \quad \text{for all } k \in [1, \dots, K]$$

where

$$\Delta \equiv \frac{1}{k-1} \log\left(\frac{R_{\mathrm{U}}}{R_{\mathrm{L}}}\right)$$

is the discrete relative return interval. We first apply this process to the root node first and then repeat the procedure for all other discrete realisations over the contract horizon. The total number of scenarios at stage t is given by $t(k-1)+1$. Notice that the adjacent nodes have $k-1$ common realisations, which significantly decreases the size of the scenario tree. Figure 14.4 illustrates a simple three-stage scenario tree with five possible realisations for each node.

A scenario, $\{s, t\}$, represents a path between time 0 and time t in the scenario tree. Denote by N_t^s a decision node associate with scenario $\{s, t\}$. Each node N_t^s has an immediate ancestor node N_{t-1}^{s-} and a set of immediate decedent nodes $N_{t+1}^{s^+} \in D(N_t^s)$. Let $p_t^s > 0$ be the probability associated with a possible realisation, r_t^s, in the scenario tree at time $t \in [0, T]$. Probability p_t^s can be calculated from the relative return probability distribution function as

$$p_t^s = \begin{cases} 1 - P(r_t \leqslant a_t^{k-1} \mid R_{t-1}^{s-}, \sigma_{t-1}^{s-}) & \text{if } s = K \\ P(a_t^{s-1} < r_t \leqslant a_t^s \mid R_{t-1}^{s-}, \sigma_{t-1}^{s-}) & \text{if } 1 < s < k \\ P(r_t \leqslant a_t^1 \mid R_{t-1}^{s-}, \sigma_{t-1}^{s-}) & \text{if } s = 1 \end{cases}$$

for a given parent node cumulative return position R_{t-1}^{s-} and tracking error decision σ_{t-1}^{s-}, where $a_t^s = \frac{1}{2}(r_t^s + r_t^{s+1})$ for all $s \in [1, \dots, k-1]$.

The scenario tree generated by the above procedure is a lattice approximation of the state space of relative returns. There are two sources of inaccuracy in using a lattice: quantisation error and specification error. Quantisation error is incurred by approximating a continuous distribution with discrete outcomes, while mismatching between the barrier level and the available lattice points causes specification error. There is a trade-off between these two types of error. As the lattice space decreases, the errors get smaller and the solution converges to the true value. However, higher accuracy requirements make the numerical calculation more cumbersome.

Our tree-generation process provides an effective approach to mitigate both the quantisation error and the specification error. Increasing the number of discrete samples does not incur too much of a calculation burden since the computational complexity is not high: the total number of nodes is $(T+1)[\frac{1}{2}T(k-1)+1]$ and the performance of the managed portfolio is usually observed quarterly or annually. We can further align a lattice level with the barrier level by adjusting the truncation error specification ε or the number of scenarios K.

We further designed a backward-induction algorithm to identify the manager's optimal tracking error policy under different scenarios. The detailed algorithm is described as follows.

Step 1. At the end of the last stage T, calculate the portfolio manager's payout $F(R_T^s)$ for each realisation r_T^s; identify the investor's knockout decision set $\{I_t^s \mid \forall t, s\}$ based on the exogenous tolerance barrier set $\{R_{t,\text{barrier}},\ t \in [1, T-1]\}$; if $I_t^s = 0$ for a certain node, calculate the manager's knockout utility based on a prespecified contract.[7]

Step 2. Let $t = T-1$, find the manager's optimal tracking error decision, σ_{T-1}^{s-}, at the beginning of the last period by solving the manager's utility maximisation problem

$$V_{T-1}^s(R_{T-1}^s) = \max_{\sigma_{T-1}^s \in [\sigma_{T-1,\mathrm{L}}, \sigma_{T-1,\mathrm{H}}]} \sum_{N_T^{s+} \in D(N_{T-1}^s)} p_T^{s+} U_T^{s+}(F(R_T^{s+}))$$

for all scenarios with $I_{T-1}^s = 1$.

Step 3. Let $t = t-1$, find the manager's optimal tracking error decision, σ_t^s, by solving the following optimisation problem

$$V_t^s(R_t^s) = \max_{\sigma_t^s \in [\sigma_{t,\mathrm{L}}, \sigma_{t,\mathrm{H}}]} \sum_{N_{t+1}^{s+} \in D(N_t^s)} p_{t+1}^{s+} V_{t+1}^{s+}$$

for all scenarios with $I_t^s = 1$ at time t. If $t = 0$, go to step 4; otherwise, repeat step 3.

Step 4. Stop and identify the optimal tracking error decision rules for the manager's utility maximisation problem.

The algorithm first calculates managers' utility at the end of contract period T based on the compensation contract and the ratio between the cumulative returns of the managed and benchmark portfolios. It then identifies the status of all the nodes at the beginning of the last stage; more specifically, if the managed portfolio has not hit the knockout barrier, an optimisation problem is solved to find the optimal tracking error which maximises the expected utility of the current node; otherwise, the manager is out. A similar procedure is then applied to the penultimate stage and so on. The manager's optimal contingent tracking error strategy is obtained by working back through all the nodes in the scenario tree. Correspondingly, the object value of the root node optimisation problem is the maximum expected utility at time T.

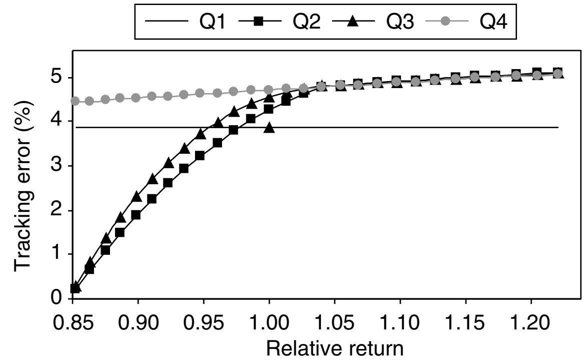
14.4 DYNAMIC DECISION MAKING UNDER VARIOUS FEE SCHEDULES

This section illustrates a manager's risk-taking behaviour in a multistage investment environment with a focus on the effects of time and performance. Our analyses include the three most popular compensation structures: proportional fees, asymmetric incentive fees and symmetric incentive fees. We find that managers show a much richer range of risk behaviours than in the single-stage model as time progresses. The manager displays a remarkable prudence as they strive to preserve their long-term franchise (the present value of future fees) in the earlier stage; however, they prefer to increase fund's risk exposure in the later stages. Our observations show that the short-sighted single-stage model exaggerates the effects of the incentive and that a well-defined knockout barrier further tempers the fund manager's risk appetite. While the presence of a knockout barrier aligns the manager's behaviour with the investor's interests, our analyses also indicate that too little tolerance of underperformance will force the manager towards a passive position while the client still pays active fees, and capping fees too early will cause a lock-in effect.

14.4.1 A proportional asset-based fee

Although performance-based compensation structures have been adopted by many investment companies, the traditional and still dominant form of compensation in the mutual fund and institutional asset management industry is an asset-based fee, where the management firm receives a fraction of total assets under management. Our analysis shows that even under such a traditional fee

Figure 14.5 Active risk decisions at different investment stages under an asset-based fee structure



This figure shows the portfolio manager's state dependent active risk decision at the beginning of each quarter. The parameters for the fraction of fund fee contract are: $R_{\text{barrier}} = 0.85$, $k = 1\%$, $\gamma = 4$, $\sigma_{\text{L}} = 0$, $\sigma_{\text{H}} = 10\%$.

structure, with no explicit performance fee, the manager still shows a certain level of risk-shifting behaviour.

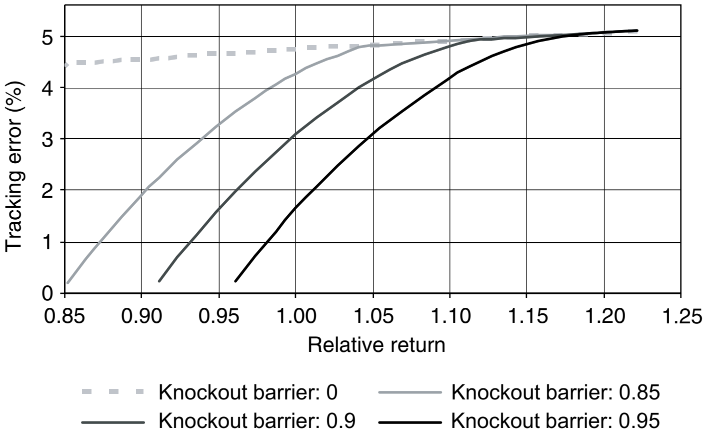
Since most investors evaluate the performance of their money manager on an annual basis,[8] the investment horizon in our analyses is set to one year with quarterly rebalance, ie, $T = 4$. We measure returns as of the start of the horizon, so that initially the manager's relative return is set equal to 1, ie, $R_0 = 1$. To enable the manager to exhibit aggressive risk-shifting behaviour, quarterly active risk is allowed to range from $\sigma_{\text{L}} = 0$ to $\sigma_{\text{U}} = 10\%$ per quarter (20% annually). The knockout barrier is set to 0.85, ie, the management contract will be immediately terminated if the cumulative realised return at the end of a quarter is 15% below the benchmark portfolio at the end of any quarter. If the manager's contract is allowed to run through the last quarter, their final payout will be 1% of the fund's wealth at the end of the year.[9] The risk free rate is assumed to be 5% per annum. We also set the portfolio manager's relative risk aversion parameter equal to 4 and the initial wealth to 1. This will be referred to as the "base case" throughout this chapter.

Figure 14.5 examines the manager's optimal active risk choice at different times. The manager's active risk exposure depends both on time and the fund's cumulative relative return. The flat risk decision rule displayed for Q1 requires some explanation. The fund's starting cumulative relative performance is fixed at 1. Hence, the optimal active risk decision is made only at the point where relative return equals 1. However, we find it helpful to represent this decision as a horizontal line, so that the decisions made at the beginnings of each of the four quarters are displayed on an equal footing. We remind the reader that, in this chart and in similar charts to follow, this decision assumes a relative return of 1 by placing a triangle on the horizontal line representing the decision rule for Q1 where the relative return is 1.

Figure 14.5 shows that the fund manager chooses a conservative active risk policy in the earlier stages. At the beginning of Q1, the manager adopts a lower active risk than in later stages at the point where the fund's relative return is 1. At the beginning of Q2, we observe the same trend across all the states. Contrary to popular belief, the manager prefers conservative investment decisions (low active risk), even when the performance is significantly behind the benchmark, as the risk of being fired due to crossing the knockout barrier hangs like the Sword of Damocles over them. As performance gets better, the manager gradually increases the fund's active risk to capture future upside returns. Dynamically this works like portfolio insurance on active returns. The manager switches dynamically between the risk-free asset (the index fund) and the risky asset (the active fund), taking higher active risk when returns are good.

As we enter later stages (Q3 and Q4), investment decisions become more aggressive. This reveals an important trade-off between the knockout penalty and the asset-based fee compensation. The manager receives a percentage of the market value of the fund's assets as compensation and is therefore inclined to take higher risks in order to increase the probability of a larger management fee. However, a higher active risk also increases the probability of being fired and thus losing all future management fee flows. On top of this, risk aversion on the manager's part and information ratio decay also decrease the marginal contribution to revenues for larger active risks. When the fund's performance trails the benchmark significantly, the probability of being fired increases quickly. Soon the knockout penalty dominates the potential upside of positive performance on asset-based fees. Eventually, the manager prefers a conservative investment strategy. As the fund's performance improves, the marginal knockout penalty decreases while the expected management fee increases and the manager tends to increase the level of active risk.

Figure 14.6 The effects of a knockout barrier on active risk decisions with a propositional asset-based fee structure



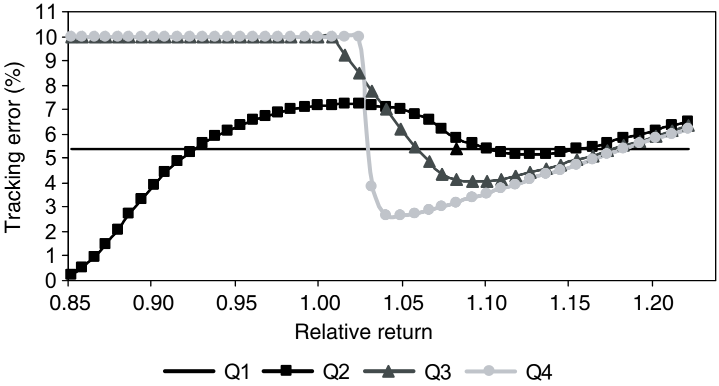
This figure shows the portfolio manager's active risk decision at the beginning of the second investment stage with three different knockout barriers: 0, 0.85, 0.9 and 0.95. All the other parameters are as in Figure 14.3.

While hitting the barrier early means losing future earnings ability and hence incurs a larger knockout cost, this is of less concern in later stages. The conservative strategy dominates in the earlier stages, especially when the fund's performance is close to the barrier.

Figure 14.6 compares the manager's active risk decisions under different knockout barrier specifications. The line with a knockout barrier of zero corresponds to the case with no barrier at all, since in our model the value of the fund is always above zero. For a given relative return, the managed fund's active risk level increases as the knockout barrier decreases. A lower knockout barrier decreases the chance of termination, which in turn results in a smaller knockout penalty cost. Note that the gaps among all lines increase, when the fund's performance deteriorates. A fund with a higher knockout barrier adopts a more conservative investment strategy because of the increased danger of the manager being fired.

Our analysis has shown that even a propositional asset-based fee may provide incentives for a fund manager to eventually take higher risks. Imposing a knockout barrier mitigates the manager's risk-shifting incentive as higher active risk leads to a larger expected knockout penalty. The multi-stage setting further illustrates that it

Figure 14.7 Active risk decisions at different investment stages under an asymmetric incentive-fee structure



This figure shows the portfolio manager's state-dependent active risk decision at the beginning of each quarter. The parameters for the asymmetric incentive compensation contract are: $R_h = 1.0$, $R_{\text{barrier}} = 0.85$, $k_1 = 1\%$, $k_2 = 20\%$, $\gamma = 4$, $T = 4$, $\sigma_L = 0$, $\sigma_H = 10\%$.

is not in the manager's best interests to adopt a myopic single-stage decision rule, since it ignores the benefit of future fee flows.

14.4.2 Asymmetric incentive fee

Asymmetric incentive-fee structures are very popular among hedge funds and commodity trading advisors (CTAs). The incentive fee is similar to a call option, with the hurdle return corresponding to the strike price. Traditional single-period analysis indicates that within this structure a fund manager is tempted to invest more aggressively, as this will increase the expected value of their performance option. Our analyses will show that risk decisions are much more complicated in a multi-stage setting. As in the base case of a proportional asset-based fee, we still assume the manager faces a constant knock-out barrier over the whole investment horizon. All other settings and parameters, except the compensation structure, are the same. We assume that the manager earns a fixed management fee of 1% of the initial assets, plus an incentive fee equal to 20% of the fund's excess return above the high watermark, which is set to 1.

Figure 14.7 shows the manager's risk rules as a function of their fund's performance status. While a manager with an asset-based compensation structure prefers to monotonically increase active risk as the fund's performance improves or as the investment horizon approaches, Figure 14.7 indicates that a hedge fund manager has a much more complex risk behaviour (as in Figure 14.5, the flat active risk observation at stage 1 is for consistency of presentation). We start with the last stage's active risk decisions as depicted by the Q4 decision rule, which in fact is a single-stage utility maximisation problem. The manager exhibits three areas of economic behaviour.

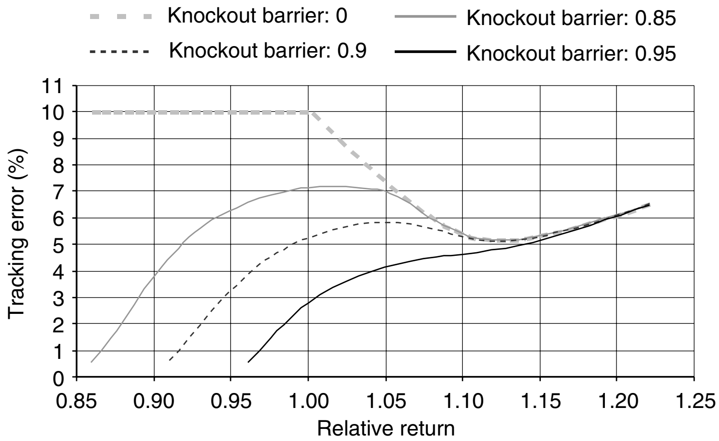
Case 1: Poor performance. When the fund's relative performance is less than 1, the manager takes the maximum allowed risk of 10%. As their performance option is out-of-the-money, it is optimal for the manager to play aggressively to increase the option value.

Case 2: Intermediary performance. As the performance option moves into-the-money, the manager tries to protect its terminal value by reducing investment risks. Figure 14.7 shows that the portfolio's active risk drops sharply from 10% to 2.64% as relative performance increases from 1.02 to 1.05. Over this region, the manager abruptly reverses their strategy from aggressive to conservative. The major motive for this change is the incentive to lock-in the realised option gain. For a lower risk aversion ($\gamma < 4$) this effect would be less pronounced.[10]

Case 3: Good performance. If the option moves further into-the-money, the manager prefers to increase active risk again. With a positive information ratio and a performance-sharing arrangement in place, there is a strong incentive to produce alpha.

Compared with the Q4 decision rule, the decisions at Q3 are qualitatively similar but less dramatic. The investment behaviours in the two investment stages are very similar when the option is either deeply in- or deeply out-of-the-money. At this stage, the manager is not as pressed for time, and this allows them to take more active risk. If risk taking pays off, the manager can lock in the gain in the last stage. If, on the other hand, the active investment suffers poor returns, the manager still has one more chance to bet in the last stage. Hence, the level of risk chosen in the decision rule for the penultimate stage is generally above that of the last stage. This effect is more significant when the relative performance falls in the region between aggressive and conservative.

Figure 14.8 The effects of a knockout barrier on active risk decisions with an asymmetric incentive-fee structure



This figure shows the portfolio manager's active risk decision at the beginning of the second investment stage with four different knockout barriers: 0, 0.85, 0.9 and 0.95. All other parameters are as in Figure 14.5.

A surprising observation from Figure 14.7 is the manager's prudent investment behaviour when their incentive option is deep out-of-the-money, as it is at the beginning of the second investment stage. This finding stands in strong contrast with the existing literature on the single-stage asymmetric incentive-fee contract. Grinblatt and Titman (1989) highlight the fact that, with limited liability, the manager has an incentive to take on a riskier portfolio than otherwise (see also Gollier *et al* 1997). Of course, this behaviour is not prudence per se. In fact, the reason the manager chooses such a low active risk is to avoid hitting the knockout barrier at an early stage, which would mean sacrificing not only the potential incentive compensation but also the present value of the future percentage fee. The option-like nature of the incentive compensation will dominate the knockout penalty as the fund's performance is far above the barrier. This leads to larger active risk as shown on the right side of the Q2 line.

Figure 14.8 illustrates the manager's active risk choices under different knockout barrier levels at the beginning of the second stage. Comparing cases with a knockout barrier to the case without, the manager shows prudent investment preference, especially when the fund is significantly trailing the benchmark. When the fund's relative performance is far ahead of the benchmark, the effects of the knockout barrier will play a minor role in determining the optimal level of investment risk. Essentially, the knockout barrier serves as a control mechanism for unobserved volatility, ie, even though the investor cannot observe volatility directly in our setting,[11] they can observe one of its consequences: unpleasantly large returns.

In summary, our analyses suggest that hedge-fund managers will follow a much more complex risk strategy in a multi-stage setting with a knockout barrier in force. More frequent observations will impose prudent behaviour, ie, to avoid risk-shifting makes business sense.

14.4.3 Symmetric compensation structure

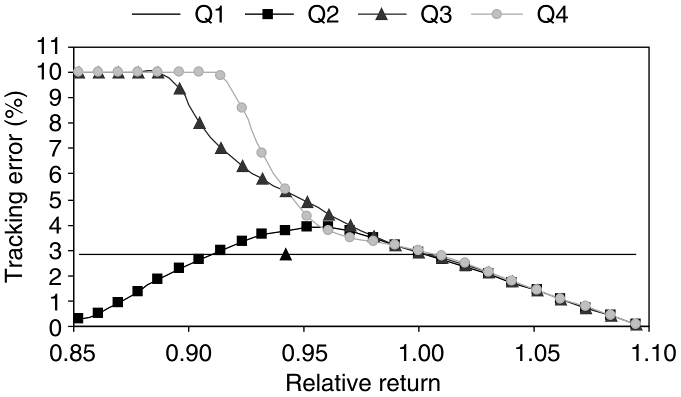
For practical purposes all fee arrangements that use a symmetric participation structure have a fixed component, as well as a variable component which is symmetric about the benchmark return. This compensation structure has recently gained popularity in the institutional and mutual fund industries. A typical fund with a fulcrum-fee structure caps the maximum negative impact of the variable fee, and this also limits the maximum attainable fee. Crucially, the floor will ensure that the manager's total compensation remains positive.

In our symmetric institutional fund example, the amount of the fixed fee is equivalent to 1% of the fund's initial wealth, and the proportional incentive part is 1% of the excess return over the benchmark. The floor and cap hurdle rates are set to 0.9 and 1.1, respectively, and the knockout barrier is set at 0.85. Note that a knockout barrier above the lower hurdle rate would make little sense, as it would make the latter irrelevant until the last period.

The manager's optimal risk strategies at different decision times through the year are shown in Figure 14.9. (Again, at stage 1 we only make a decision at the point where relative return equals 1.) Based on the manager's active or passive risk-taking style, the fund's relative return space can be divided into two regions: the passive, benchmark-tracking region and the active risk management region.

The passive benchmark tracking region is located on the right side of the cap relative return point, which is $R_C = 1.1$ in these examples.

Figure 14.9 Active risk decisions at different investment stages under a fulcrum-fee structure



This figure shows the portfolio manager's optimal active risk decision as a function of relative return at the beginning of each quarter. The parameters for the fulcrum fee are: $R_C = 1.1$, $R_F = 0.9$, $k_1 = 1\%$, $k_2 = 1\%$, $R_{\text{barrier}} = 0.85$, $\gamma = 4$, $T = 4$, $\sigma_L = 0$, $\sigma_H = 10\%$.

The manager's optimal policy is to hold the benchmark portfolio and lock in the gains. The explanation of the passive benchmark tracking region is very intuitive. When the manager performs very well over the earlier investment stages, they are near or above the concave portion of the payout schedule. Therefore, the best strategy is to closely track the benchmark ("closet indexing") and lock in the realised gain. Following a passive index strategy, however, is not necessarily in the best interests of the investor, who is paying active management fees in the expectation of receiving a positive alpha. Even though the manager has achieved superior performance in an earlier stage, it would be beneficial to the investor if the manager could maintain the level of active risk for the rest of the investment horizon. Otherwise the breadth of an investment strategy clearly suffers. To avoid a gain-lock-in effect, the ceiling of the performance-based fee should in generally not be too low. Good pragmatic advice is for the investor to set it where they think luck begins.

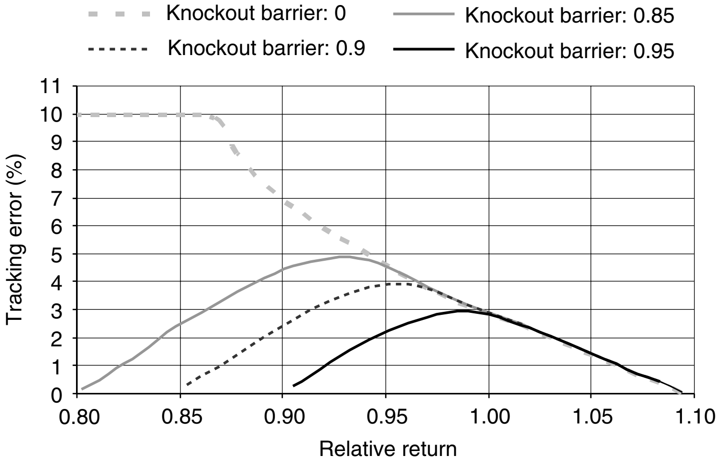
The active risk management region is located on the left side of the cap rate point. Over this region, the manager's decision is similar to the asymmetric case except when the fund's performance is significantly above the benchmark. The optimal investment policy depends on both the number of investment stages remaining and the fund's cumulative performance relative to the benchmark. To highlight the time effect, we discuss the earlier stages and later stages separately.

Case 1: Earlier stages. The Q2 line shows the manager's optimal active risk policy at the beginning of the second investment period. The decision curve is humped because the knockout threat (loss of future fee opportunities) overpowers the incentive to increase the value of the implicit long call on out-performance in the region when the fund significantly underperforms its benchmark. As performance improves and compensation becomes more symmetric the manager gradually increases active risk in order to maximise utility. As performance approaches the determined performance cap, risk-taking only offers downside possibilities for the manager (but not for the client) and we see the familiar reigning in of active risk. Compared with the manager's active risk decision at stage 2 at the point where the portfolio's relative return is 1, the manager's decision at stage 1 is more conservative because the knockout penalty in stage 1 is larger than that in stage 2.

Case 2: Later stages. The logic changes slightly in later stages, Q3 and Q4, when underperformance leads the manager to expect only the minimum (floor) fee. Now the threat of termination largely loses its disciplining influence, while the all-or-nothing feature of owning a call on out-performance makes the policy of maximising active risk optimal. It is all the same to the manager whether they are terminated as a result of hitting the knockout barrier or lose their incentive fee because of bad performance. The level of active risk decreases almost monotonically as the fund's performance improves, because of the concave utility (large fee payments carry less utility) and the capped upside compensation (fee payments do not become very large anyway).

Compared with the optimal policy of the asymmetric compensation structure shown in Figure 14.7, the manager follows similar active strategies at the beginning of the third and the fourth investment stages in the middle region (between cap and floor). However, the gap between the decision curves of these two stages becomes smaller. This can be explained by the concave nature of the collar option in the neighbourhood of the cap, since limited cap compensation dampens the manager's anticipated compensation at an even earlier stage.

Figure 14.10 The effects of a knockout barrier on active risk decisions with a fulcrum-fee structure



This figure shows the portfolio manager's active risk decision at the beginning of the second investment stage with four different knockout barriers: 0, 0.8, 0.85, 0.9. All other parameters are as in Figure 14.7.

Overall, our observations confirm the manager's tendency to balance the knockout penalty and the option benefit. The manager optimally trades off termination risk against the anticipated fees. As relative performance moves away from the barrier, the marginal loss (adding one unit of risk) from termination risk decreases. Risk taking, however, has an upper limit as both concave utility and concave compensation eventually make risk-taking less attractive.

As in the previous sections, we present a comparative analysis for different knockout barriers in Figure 14.10. Intuitively, for a given performance state, a higher barrier level increases the chance of knockout and the marginal penalty cost for an additional unit of risk. Therefore, the existence of the knockout barrier mitigates the manager's risk taking incentives.

Just as too low a cap fee can push the manager to follow a passive index strategy, too low a tolerance of downside risk also provides an undesirable incentive: a passive index tracking strategy in earlier stages and an aggressive gamble at the last stage. If the earlier stage's poor performance brings the fund close to the knockout barrier, the manager has no room to take active risk. Instead, they will follow a strict indexing strategy, because even a moderate active risk exposure may leads to a high chance of being fired (up to 50% in the case of the normal distribution).

14.5 CONCLUSIONS

This chapter investigated the determination of the optimal active risk policy as a function of time and relative performance (state variables) using stochastic programming techniques.

In contrast to previous research, we explicitly modelled the possibility that a management contract will be terminated if performance drops below a minimum acceptable target level. We find that this knockout barrier will encourage the manager to invest more conservatively. This result is explained by the trade-off between the knockout penalty and the compensation incentive. Managers take little risk when the fund's performance is close to a knockout barrier, because of the large penalty cost of being fired. Essentially the knockout barrier serves as a control mechanism for unobserved volatility, ie, even though volatility, in our setting, cannot be observed directly by the investor, we can observe one of its consequences: unpleasantly large returns.

The manager carefully manoeuvres the fund to avoid hitting the barrier and therefore reduces the chance of being fired in an early stage and so losing the present value of future fees. In summary, our analyses suggest that hedge fund managers will follow a much more complex risk strategy in a multi-stage setting with a knockout barrier in effect. More frequent observations will impose prudent behaviour, ie, avoiding risk-shifting makes business sense.

Our results have important implications for active management, and are considerably different for each fee structure. At a minimum, this research will help managers and investors alike to write a more optimal compensation contract. This not only applies to the relationship between asset management firms and their clients but also to the compensation of portfolio managers within investment houses.

[1] We ignore the case of flat fees, which are independent of both asset and performance, as this does not appear to generate any interesting risk-shifting behaviour.

[2] It is beneficial to elaborate on this point a bit further. Practitioners seem to prefer asset-based fees over fixed fees despite the fact that asset-based fees make them share a client's benchmark risk. They do this because their intuition tells them that asset-based fees benefit from the underlying trend in (equity) markets. However, this intuition is wrong. Asset-based fees are

among the most primitive of derivatives and hence the price of a derivative is independent of its (real world) drift rate.

3 For a single-period model it is often convenient to calculate arithmetic outperformance. However, to be consistent with our multi-period model in the rest of this chapter, we use geometric outperformance instead.

4 We assume throughout that the manager does no market timing, so that residual risk and active risk are equal.

5 We assume that the knockout barrier is exogenously determined by the investors. It is natural to allow the barrier to be time dependent since the variance of the fund's cumulative relative performance will increase as time progresses.

6 In order to capture alpha we need to deviate from a pure contingent claims approach. Remember that option pricing does not know the concept of alpha as it does not need fair valuation. In fact option pricing is relative pricing, ie, all claims are priced correctly relative to each other, but every single claim could be far from its equilibrium value.

7 We assume the investor will pay a pro-rated portion of the fixed management fee which will be invested into a risk free account by the manager. The utility of a knockout realisation is then calculated based on the future value of the pro-rated portion of the fixed fee at time T.

8 A one-year evaluation period is considered as the most commonly used evaluation period for money manager in industry.

9 In practice, the asset-based fee is charged as a percentage of average assets under management over an investment period. For simplicity, we assume the fee is calculated based on the end-of-year asset only.

10 The authors can provide analysis to support this on request.

11 Of course, there would be no need for all this if investors could perfectly monitor manager outcome, ie, if there were neither hidden action nor hidden information.

REFERENCES

Basak, S., A. Pavlova and A. Shapiro, 2007, "Optimal Asset Allocation and Risk Shifting in Money Management", *Review of Financial Studies* 20(5), pp. 1583–1621.

Brown, K., W. Harlow and L. Starks, 1996, "Of Tournaments and Temptations: An Analysis of Managerial Incentives in the Mutual Fund Industry", *Journal of Finance* 51, pp. 85–110.

Carpenter, J., 2000, "Does Option Compensation Increase Managerial Risk Appetite?", *Journal of Finance* 55, pp. 2311–31.

Chevalier, J., and G. Ellison, 1997, "Risk Taking by Mutual Funds as a Response to Incentives", *Journal of Political Economy* 105, pp. 1167–1200.

Das, S., and R. Sundaram, 1999, "Fee Speech: Adverse Selection and the Regulation of Mutual Funds", Working Paper, Center for Law and Business, New York University.

Ferguson, R., and D. Leistikow, 2003, "Long-Run Investment Management Fee Incentives and Discriminating between Talented and Untalented Managers", *Journal of Investment Management* 4(1), pp. 47–72.

Goetzmann, W., J. Ingersoll and S. Ross, 2003, "High-Water Marks and Hedge Fund Management Contracts", *Journal of Finance* 58, pp. 1685–717.

Gollier, C., P. Koehl and J. Rochet, 1997, "Risk-Taking Behavior with Limited Liability and Risk Aversion", *Journal of Risk and Insurance* 64(2), pp. 347–70.

Grinold, R., and R. Kahn, 2000, "The Efficiency Gains of Long-Short Investing", *Financial Analysts Journal* 56(6), pp. 40–53.

Grinblatt, M., and S. Titman, 1989, "Adverse Risk Incentives and the Design of Performance-Based Contracts", *Management Science* 35(7), pp. 807–22.

Heinkel, R., and N. Stoughton, 1994, "The Dynamic of Portfolio Management Contracts", *The Review of Financial Studies* 7(2), pp. 351–87.

Hodder, J., and J. C. Jackwerth, 2007, "Incentive Contracts and Hedge Fund Management", *Journal of Financial and Quantitative Analysis*, 42(4), pp. 811–26.

Ippolito, R., 1992, "Consumer Reaction to Measures of Poor Quality: Evidence from the Mutual Fund Industry", *Journal of Law and Economics* 35, pp. 45–70.

Khorana, A., 1996, "Top Management Turnover: An Emperical Investigation of Mutual Fund Managers", *Journal of Financial Economics* 40, pp. 403–27.

Stremme, A., 2001, "Optimal Compensation for Fund Managers of Uncertain Type: The Information Advantages of Bonus Schemes", Working Paper, Department of Finance, New York University.

15

Long-Term Portfolio Choice

One of the basic questions in long-run portfolio choice is whether investors with long time horizons should hold a larger fraction of stocks in their portfolio. Academically, there is a sophisticated theoretical and empirical literature developing around this very question that takes predictability, learning, human capital and parameter uncertainty into account with very different results on how the time horizon might affect asset allocation.[1] Contrary to academia, the asset management industry have already made their mind up with an unconditional "yes". Investors with longer time horizons should hold more equity. This chapter reviews the basic techniques and literature employed in long-term portfolio choice.

15.1 LONG-TERM PORTFOLIO CHOICE UNDER IID ASSUMPTIONS: TWO FALLACIES

We start with in a world with independent and identically distributed (iid) returns (there is no exploitable pattern in returns). In other words, investment opportunities are constant. With no time variation in investment opportunities, investors have no reason to hedge against changes in future investment opportunities. This is sometimes overlooked by practitioners. Here are two of the most common fallacies.

15.1.1 Time diversification

The most common argument is that time somewhat diversifies risk. It is argued that investing US$100 in an asset with 2% expected return and 10% risk per quarter is more risky than investing US$100 into this asset for 400 quarters. This is because the average return of 2% becomes increasingly safe as the standard deviation of average returns after 400 quarters shrinks to $0.1/\sqrt{400} = 0.005$. However, this is an inappropriate application of the law of large numbers. Diversification works because risks are subdivided, ie, splitting US$100 into 100 uncorrelated assets with US$1 weight each will reduce risk. Adding 100 positions of US$100 on top of each other will increase risk. Long-term investing is the equivalent of repeatedly investing US$100 (plus/minus profits). Another way to see this is to apply the "Chain Store Paradox" to investing. Suppose you are at the end of quarter 399, that is, you have one quarter left. Because you are then a one-period investor you will not accept this one-period bet. So quarter 399 becomes your last quarter. At the end of quarter 398 you again have only one quarter left (as you do not invest in the last quarter) and will reject. This unravels until the first quarter where again you will not invest.[2]

15.1.2 Shortfall risk

The investment management industry usually works with risk return diagrams where risk is measured as standard deviation. Under iid assumptions, risk grows with the standard deviation of time while (log) returns grow with time. The longer the time horizon, the larger the Sharpe ratio. Of course, this does not mean that stocks are more attractive for longer time horizons. In fact, all assets would look more attractive for longer time horizons. Sharpe ratios must only be used to compare assets for a given time horizon, not to compare one asset across time. This is directly related to the use of shortfall risk. Sharpe ratios measure the likelihood of underperforming cash. A Sharpe ratio of 1.96 means there is only a 2.5% likelihood of returning less than cash. It is well established that shortfall risks are not a valid risk measure as they assume risk-loving behaviour in the tail of a distribution, which is the opposite of all we know about investor behaviour. While the likelihood of a shortfall will decrease with the investment horizon, the maximum loss will increase. For risk-averse investors this will offset the appeal of lower shortfall risks.

15.2 PREDICTABILITY AND THE TERM STRUCTURE OF RISK

The academic literature has shown a great revival of interest in dynamic asset allocation and long-term portfolio choice since the publication of Campbell and Viceira (2002). Their work has been driven by a growing body of empirical evidence that equity excess returns can be predicted by dividend yields, inflation rates or interest rates that might very well proxy for economic state variables. Similar evidence is found for bond markets, where the term spread is known as a predictor of excess bond returns.[3] In short, the literature suggests that expected returns contain a time-varying component that drives the predictability of returns. The high degree of persistence in explanatory variables creates a predictive component that is stronger over long horizons than over short horizons. The dividend–price ratio, the earnings–price ratio and the book-to-market ratio are usually quoted, while other variables have also found to be important (see, for example, Lettau and Ludvigson 2001; Menzley *et al* 2004).

It is this predictability (or, better, the existence of time-varying investment opportunities) that creates inter-temporal hedging demand in the spirit of Merton (1971). Investors want to buy assets that perform well if future investment opportunities deteriorate. A second consequence of predictability is the existence of time-horizon effects. If returns can be shown to be mean reverting (equity losses today translate into higher expected returns tomorrow), they might be less risky in the long run than under the iid assumption.

To illustrate the effects of predictability on the term structure of risk and correlation (how risks change when the time horizon lengthens) we review the basic set-up of the Campbell and Viceira model. Let $x_{t+1} = r_{t+1} - r_{0,t+1}\iota$ denote a $k_1 \times 1$ vector of excess returns, ie, a $k_1 \times 1$ vector of asset returns r_{t+1} over a risky benchmark return $r_{0,t+1}$, where ι denotes a $k_1 \times 1$ vector of 1s. We can view x_{t+1} as excess returns on a set of core asset classes. Often the real return on cash is used as a risky benchmark return for long-term investors. Also define an $m \times 1$ vector of economic state variables s_t that are chosen on the basis of their ability to forecast future excess returns. Stack all these variables into a single $(1 + k_1 + m) \times 1$ vector

$$y_{t+1} = \begin{bmatrix} r_{0,t+1} \\ x_{t+1} - r_{0,t+1}\iota \\ s_t \end{bmatrix} \tag{15.1}$$

Next we formulate a first-order vector autoregressive (VAR) process to model the asset return dynamics that are needed to create a meaningful long-term asset-allocation framework

$$y_{t+1} = a + By_t + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim N(0, \Omega_{\varepsilon\varepsilon}) \tag{15.2}$$

To further generalise our notation, we assume a second set of shorter time-series alternative asset returns (exogenous returns, ie, their lagged values have no explanatory power for either state variables or other asset returns) in a $k_2 \times 1$ vector $x_{2,t+1}$

$$x_{2,t+1} = c + \mathit{\Psi}_0 y_{t+1} + \mathit{\Psi}_1 y_t + \mathit{\Pi} x_{2,t} + \nu_{t+1}, \quad \nu_{t+1} \sim N(0, \mathit{\Omega}_{vv}) \tag{15.3}$$

Here $\mathit{\Psi}_0, \mathit{\Psi}_1$ are regression $k_2 \times (1+k_1+m)$ coefficient matrixes (simply estimated by equation-by-equation OLS) and

$$\mathit{\Pi} = \operatorname{diag}(\vartheta_1, \vartheta_2, \dots, \vartheta_{k_2})$$

denotes a diagonal $k_1 \times k_1$ matrix, ie, past lags of $x_{2,t+1}$ only affect their own history and nothing else

$$z_{t+1} = \begin{bmatrix} y_{t+1} \\ x_{2,t+1} \end{bmatrix} = \mathit{\Phi}_0 + \mathit{\Phi}_1 z_t + u_{t+1}, \quad u_{t+1} \sim N(0, \mathit{\Omega}) \tag{15.4}$$

$$\left.\begin{aligned} \mathit{\Phi}_0 &= \begin{bmatrix} a \\ c + \mathit{\Psi}_0 a \end{bmatrix} \\ \mathit{\Phi}_1 &= \begin{bmatrix} B & 0 \\ \mathit{\Psi}_1 + \mathit{\Psi}_0 B & \mathit{\Pi} \end{bmatrix} \\ \mathit{\Omega} &= \begin{bmatrix} \mathit{\Omega}_{ee} & \mathit{\Omega}_{ee}\mathit{\Psi}_0' \\ \mathit{\Psi}_0\mathit{\Omega}_{ee} & \mathit{\Omega}_{vv} + \mathit{\Psi}_0\mathit{\Omega}_{ee}\mathit{\Psi}' \end{bmatrix} \end{aligned}\right\} \tag{15.5}$$

The idea here is that only unexpected variations (those not explained by state variables) pose a risk. We can now fix risk and return at various time horizons by substituting Equation 15.4 forward

$$z_{t+j} = \sum_{i=1}^{j-1} \mathit{\Phi}_1^i \mathit{\Phi}_0 + \mathit{\Phi}_1^j z_t + \sum_{i=0}^{j-1} \mathit{\Phi}_1^i u_{t+j-1} \tag{15.6}$$

Summing over n periods (to arrive at cumulative returns) and applying both the expectations as well as the variance operator leaves us with expressions for multi-period risk and return

$$\mu^{(n)} = \frac{1}{n} E\biggl(\sum_{j=1}^{n} z_{t+j}\biggr) = \frac{1}{n} \sum_{j=1}^{n} \biggl(\sum_{i=1}^{j-1} \mathit{\Phi}_1^i \mathit{\Phi}_0 + \mathit{\Phi}_1^j z_t\biggr) \tag{15.7}$$

$$\mathit{\Omega}^{(n)} = \frac{1}{n} \sum_{j=1}^{n} \biggl(\biggl(\sum_{i=0}^{j-1} \mathit{\Phi}_1^i\biggr) \mathit{\Omega} \biggl(\sum_{i=0}^{j-1} \mathit{\Phi}_1^i\biggr)'\biggr) \tag{15.8}$$

Note that Equations 15.7 and 15.8 contain risks and returns for risky benchmark return, core assets excess return, alternative assets excess

Table 15.1 VAR coefficients from Campbell and Viceira (2002, p. 105)

$\hat{B}$	Real	Equities	Bonds	Nominal	Dividend	Term
from	T-bill	(excess)	(excess)	T-bill	yield	spread
Eqn 15.2	rtb$_t$	xr$_t$	xb$_t$	y_t	$(d-p)_t$	spr$_t$
rtb$_{t+1}$	0.3055	−0.0543	0.1185	0.6773	−0.006	−0.808
xr$_{t+1}$	0.1107	0.0795	−0.0823	−0.0205	0.1351	1.3653
xb$_{t+1}$	0.2007	0.1061	−0.1974	−0.1153	0.0122	2.623
y_{t+1}	−0.042	−0.0123	0.0365	0.9196	−0.0052	−0.0189
$(d-p)_{t+1}$	−0.562	−0.1291	0.3479	−0.6522	0.8374	−1.7388
spr$_{t+1}$	0.0195	0.0023	−0.0127	0.0872	0.004	0.8227

Each row denotes one equation within a standard VAR(1) model, ie, the first row represents a regression of real T-bill rates against lagged real T-bills, lagged equity excess returns, lagged bond excess returns, etc.

Table 15.2 Residual covariance matrix from Campbell and Viceira (2002, p. 105)

$\Omega_{\varepsilon\varepsilon}$ from						
Eqn 15.2	rtb$_t$	xr$_t$	xb$_t$	y_t	$(d-p)_t$	spr$_t$
rtb$_t$	58.5531	−25.365	−0.54432	1.162339	15.65108	−1.25854
xr$_t$	−25.365	310.8874	−2.2397	−3.11028	−209.799	3.435084
xb$_t$	−0.54432	−2.2397	25.81656	−4.02085	−3.95903	1.288343
y_t	1.162339	−3.11028	−4.02085	1.500625	3.73846	−1.07335
$(d-p)_t$	15.65108	−209.799	−3.95903	3.73846	263.5103	−2.94004
spr$_t$	−1.25854	3.435084	1.288343	−1.07335	−2.94004	0.958441

return as well as state variables. We still need to address the right components with the help of a pick matrix. Suppose we model a six-variable VAR(1) as in Campbell and Viceira (2002) for real T-bills (risky benchmark return), excess stock and bond returns, nominal T-bills, dividend yield and term spread in exactly this ordering. The VAR(1) coefficients and residual covariance matrix are given in Tables 15.1 and 15.2.

In this case $\Omega^{(n)}$ is a 6 × 6 matrix and calculated from

$$\Omega^{(n)} = \frac{1}{n}\sum_{j=1}^{n}\left(\left(\sum_{i=0}^{j-1}\hat{B}^{i}\right)\hat{\Omega}_{\varepsilon\varepsilon}\left(\sum_{i=0}^{j-1}\hat{B}^{i}\right)'\right)$$

Suppose we want to learn the term structure of risk for total equity returns. We cannot directly read this from $\Omega^{(n)}$. Instead, we need to create a vector that extracts the relevant information. Setting $\delta_{\text{eq}} = (1,1,0,0,0,0)'$, we can model the term structure of risk for equities as

$$\sigma^{(n)}_{\text{equities}} = \delta^{\text{T}}_{\text{eq}}\Omega^{(n)}\delta_{\text{eq}} \tag{15.9}$$

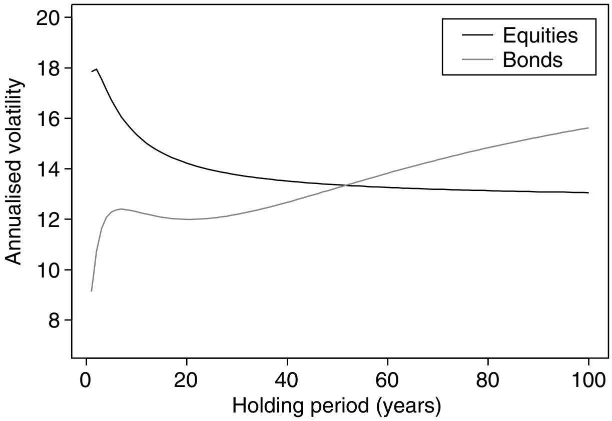
We can be creative with this. Assume we need to calculate the term structure of inflation risk even though it is not directly included in our VAR. As inflation amounts to the spread between nominal returns and real returns we can extract inflation risk using $\delta_{\text{infl}} = (-1,0,0,1,0,0)'$ instead. Let us first plot the term structure of risk for bonds and equities in Figure 15.1.

We plot the conditional standard deviation, ie, the standard deviation after predictable movements have been taken out (predictable movements do not represent risks). Equities are mean reverting (risk is decreasing as the horizon lengthens), while bonds are mean averting (bonds become riskier in the long run). Interestingly, this was picked up quickly by pension funds that viewed this as a way to defend their often unreasonably large equity allocations.[4]

The previous approach is not limited to calculating horizon-dependent volatilities. We could also use it to estimate the long-term correlation between asset returns and inflation. Let us look at the case of bonds. We need to calculate the volatility of bond returns (see above), the volatility of inflation (see above) and the covariance between bond returns and inflation. The only missing ingredient is

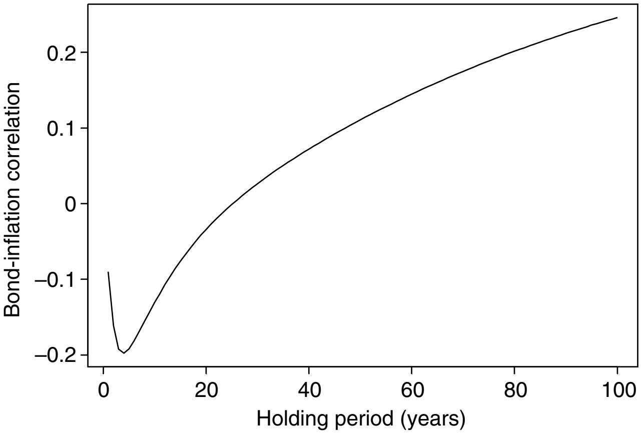
$$\begin{aligned}&\operatorname{cov}(rb_t - y_t, y_t + xb_t)^{(n)}\\&\quad = \operatorname{cov}(rb_t, y_t)^{(n)} + \operatorname{cov}(rb_t, xb_t)^{(n)} - \operatorname{cov}(y_t, y_t)^{(n)} - \operatorname{cov}(y_t, xb_t)^{(n)}\end{aligned}$$

Figure 15.1 Term structure of risk for VAR(1) for the Campbell and Viceira (2002) example



Equities are mean reverting (risk is decreasing as the horizon lengthens), while bonds are mean averting (bonds become riskier in the long run).

Figure 15.2 Correlation between bond returns and inflation



For each holding period (in years) we calculate the correlation between bond returns and inflation implied from a VAR(1) model.

from which we calculate

$$\rho = \frac{\operatorname{cov}(rb_t - y_t, y_t + xb_t)^{(n)}}{\sqrt{\operatorname{var}(rb_t - y_t)^{(n)}}\sqrt{\operatorname{var}(y_t + xb_t)^{(n)}}} \tag{15.10}$$

The results are shown in Figure 15.2.

Bond returns tend to fall when inflation increases. This effect is the strongest for the first five years. After that the impact of inflation vanishes to eventually become positive.

15.3 BAYESIAN ESTIMATES OF THE RETURN-GENERATING PROCESS

Portfolios based on risk and return estimates according to Equations 15.7 and 15.8 often show remarkable variation in portfolio weights over time depending on the parameter estimates in Equation 15.4. This arises from our departure from the assumption of independently distributed returns over time in Equations 15.2 and 15.3. However, some of the variation might be spurious and due to estimation error in Ψ_0, Ψ_1 and Ω. This is usually addressed with a flat uniform (uninformative) prior. We first stack all observations in Equation 15.2 to arrive at

$$Y = \tilde{X}\tilde{B}' + E \tag{15.11}$$

Here Y is a $T \times (1 + k_1 + m)$ matrix of observations on y_{t+1}, $\tilde{X} = (\iota \quad X)$ with X a $T \times (1 + k_1 + m)$ matrix of observations on y_t and $\tilde{B} = (c \quad B)$. The uniform prior chosen has the form

$$p(c, B, \Omega_{ee}) \propto |\Omega_{ee}|^{-(k_1+m+2)/2} I(B) \tag{15.12}$$

The indicator function $I(B)$ restricts B such that its maximum eigenvalue is less than 1. As with uninformative priors in general, coefficient estimates remain unchanged. Analytic solutions are not available, due to the discontinuity introduced by $I(B)$. Bayesians have used Gibbs sampling instead.[5] We use the following steps.

Step 1. Estimate Equation 15.11 using OLS and save E_1.

Step 2. Draw $\Omega_{\varepsilon\varepsilon,1}$ from an inverted Wishart

$$p(\Omega_{\varepsilon\varepsilon} \mid \tilde{B}, Y, X) \sim \mathrm{IW}(E_1' E_1, T)$$

Step 3. Draw the elements of $\tilde{B}_1$ from

$$N(\mathrm{vec}(\hat{\tilde{B}}_{\mathrm{OLS}}), \Omega_{\varepsilon\varepsilon,1} \otimes (\tilde{X}'\tilde{X})^{-1})$$

and reject the draw if the coefficients would lead to a non-stationary VAR.

Step 4. Take $\tilde{B}_1$ and recalculate the residuals in Equation 15.11.

Step 5. Repeat Steps 2–4 many (eg, 100,000) times. This leaves us with 100,000 pairs of $\tilde{B}_1, \tilde{\Omega}_{\varepsilon\varepsilon,1}, \dots, \tilde{B}_{100,000}, \tilde{\Omega}_{\varepsilon\varepsilon,100,000}$

Step 6. For each pair we simulate two paths (antithetic random variables) of returns for n periods in the future.

This provides us with multi-period return distributions which in turn can be used for portfolio choice.

15.4 MULTI-PERIOD PORTFOLIO CHOICE

15.4.1 Asset-only investor

Campbell and Viceira (2002) provide us with an explicit asset-allocation framework that ties investment decisions to both the conditional risk and return expectations at any point in time, as well as the decision maker's time horizon. The following exposition follows their work with some minor variations, eg, the introduction of liabilities and non-tradable real wealth.

We start with the log-linearisation of portfolio returns. This is needed because log returns are normally distributed and additive across time but not additive across assets. At the same time, simple returns are additive across assets but not across time. Neither are they normal. The log-linear portfolio return is given by

$$\begin{aligned}\underbrace{r_{a,t+1}}_{\substack{\text{Log return}\\ \text{on portfolio}}} &\approx r_{0,t+1} + w_t^{\mathrm{T}}(r_{t+1} + \tfrac{1}{2}\sigma_a^2) - \tfrac{1}{2}w_t^{\mathrm{T}}\Omega_{aa}w_t \\ &= r_{0,t+1} + \underbrace{w_t^{\mathrm{T}} r_{t+1}}_{\substack{\text{Linear combination}\\ \text{of log returns}}} + \underbrace{\tfrac{1}{2}w_t^{\mathrm{T}}(\sigma_a^2 - \Omega_{aa}w_t)}_{\text{Adjustment term}} \qquad (15.13)\end{aligned}$$

The first term in Equation 15.13 reflects the log T-bill rate. The second term calculates the arithmetic portfolio excess return by adding one-half of the individual asset variance (σ_a^2 is a vector of asset-specific variances, ie, $\sigma_a^2 = \mathrm{diag}(\Omega_{aa})$) to asset-specific log returns, r_{t+1}. However, we also need to subtract the variance drain on log portfolio returns by subtracting one-half of the portfolio variance. The log return on a portfolio is not the same as the weighted average of individual asset log returns. We need to correct this error with $\frac{1}{2}w_t^{\mathrm{T}}(\sigma_a^2 - \Omega_{aa}w_t)$. This, however, is only an approximation that will become worse as the time horizon lengthens. Essentially, this approximation is only needed to satisfy the academic obsession with closed-form solutions. Alternatively, we could simulate period-n returns using a parametric (simulated from covariance matrix of residuals) or semi-parametric VAR (bootstrapped from empirical residuals) to simulate a scenario of returns and optimise the utility function of choice. While the author has not compared both methods, he conjectures that the latter (and definitely correct) solution will provide different results.

Given that log returns are additive over time we arrive at an expression for multi-period log returns

$$\begin{aligned} r_{a,t+n}^{(n)} &= \sum_{i=1}^{n} r_{a,t+i} \\ &= \sum_{i=1}^{n} (r_{0,t+i} + w_t^{\mathrm{T}}(r_{t+i} + \tfrac{1}{2}\sigma_a^2) - \tfrac{1}{2} w_t^{\mathrm{T}} \varOmega_{aa} w_t) \\ &= r_{0,t+n}^{(n)} + w_t^{\mathrm{T}}(r_{t+n}^{(n)} + \tfrac{1}{2} n\sigma_a^2) - \tfrac{1}{2} n w_t^{\mathrm{T}} \varOmega_{aa} w_t \end{aligned} \tag{15.14}$$

Note that

$$\sum_{i=1}^{n} r_{0,t+i} = r_{0,t+n}^{(n)} \quad \text{and} \quad \sum_{i=1}^{n} r_{t+i} = r_{t+n}^{(n)}$$

Taking the expectations operator on Equation 15.14 defining $Er_{t+n}^{(n)} = n\mu^{(n)}$ and focusing on terms that are linked to portfolio weights (our decision variable), we get

$$\begin{aligned} E(r_{a,t+n}^{(n)}) &= w_t^{\mathrm{T}}(n\mu^{(n)} + \tfrac{1}{2} n\sigma_a^2) - \tfrac{1}{2} n w_t^{\mathrm{T}} \varOmega_{aa} w_t \\ &= n[w_t^{\mathrm{T}}(\mu^{(n)} + \tfrac{1}{2}\sigma_a^2) - \tfrac{1}{2} w_t^{\mathrm{T}} \varOmega_{aa} w_t] \end{aligned} \tag{15.15}$$

We repeat this exercise, applying the variance operator on Equation 15.14 and defining

$$\begin{aligned} \operatorname{var}(r_{0,t+n}^{(n)}) &= n\sigma_c^{(n)2} \\ \operatorname{cov}(r_{0,t+n}^{(n)}, r_{a,t+n}^{(n)}) &= n\sigma_{ac}^{(n)} \\ \operatorname{var}(w_t^{\mathrm{T}} r_{t+n}^{(n)}) &= n w_t^{\mathrm{T}} \varOmega_{aa}^{(n)} w_t \end{aligned}$$

we have

$$\operatorname{var}(r_{a,t+n}^{(n)}) = n w_t^{\mathrm{T}} \varOmega_{aa}^{(n)} w_t + 2n w_t^{\mathrm{T}} \sigma_{ac}^{(n)2} \tag{15.16}$$

Assuming CRRA preferences with risk aversion, γ, we arrive at the following portfolio optimisation problem

$$w_{\text{Asset only}}^{(n)} = \arg\max E(r_{a,t+n}^{(n)}) - \tfrac{1}{2}(1-\gamma)\operatorname{var}(r_{a,t+n}^{(n)}) \tag{15.17}$$

Taking derivatives with respect to w_t yields the optimal solution

$$w^{(n)}_{\text{Asset only}} = \frac{1}{\gamma}\left[\left(1-\frac{1}{\gamma}\right)\Omega^{(n)}_{aa} + \frac{1}{\gamma}\Omega_{aa}\right]^{-1}[\mu^{(n)} + \tfrac{1}{2}\sigma^2_a + (1-\gamma)\sigma^{(n)}_{ac}] \tag{15.18}$$

The optimal portfolio consists of a speculative and hedging demand (two-fund separation)

$$w^{(n)}_{\text{Asset only, speculative}} = \frac{1}{\gamma}\left[\left(1-\frac{1}{\gamma}\right)\Omega^{(n)}_{aa} + \frac{1}{\gamma}\Omega_{aa}\right]^{-1}[\mu^{(n)} + \tfrac{1}{2}\sigma^2_a] \tag{15.19}$$

$$w^{(n)}_{\text{Asset only, hedging}} = \frac{1-\gamma}{\gamma}\left[\left(1-\frac{1}{\gamma}\right)\Omega^{(n)}_{aa} + \frac{1}{\gamma}\Omega_{aa}\right]^{-1}\sigma^{(n)}_{ac} \tag{15.20}$$

For highly risk-averse investors $\gamma \to \infty$, we get

$$w^{(n)}_{\text{Asset only, hedging}} = -[\Omega^{(n)}_{aa}]^{-1}\sigma^{(n)}_{ac}$$

of hedging demand and zero speculative demand. Assets that perform well when T-bills are falling (and hence reinvestment rates get smaller) are in positive demand.

15.4.2 Benchmark-relative investor

Let us extend this framework by incorporating a benchmark portfolio. This benchmark portfolio could be pension liabilities, inflation linked bonds or any other investable asset. Now, Equation 15.13 becomes

$$r_{\text{Active},\,t+1} = w^{\mathrm{T}}_t(r_{t+1} + \tfrac{1}{2}\sigma^2_a) - \tfrac{1}{2}w^{\mathrm{T}}_t\Omega_{aa}w - r_{b,t+1} \tag{15.21}$$

where $r_{b,t+1}$ denotes the log excess (over T-bills) returns for a benchmark portfolio. The return on T-bills drops out, as it is contained in both the long side (assets) and the short side (benchmark portfolio). From here on it is straightforward to arrive at expressions for multi-period risk and return

$$E(r^{(n)}_{\text{Active},\,t+n}) = n[w^{\mathrm{T}}_t(\mu^{(n)} + \tfrac{1}{2}\sigma^2_a) - \tfrac{1}{2}w^{\mathrm{T}}_t\Omega_{aa}w_t - \mu^{(n)}_b] \tag{15.22}$$

$$\operatorname{var}(r^{(n)}_{\text{Active},\,t+n}) = n[w^{\mathrm{T}}_t\Omega^{(n)}_{aa}w_t - 2nw^{\mathrm{T}}_t\sigma^{(n)2}_{ab}] \tag{15.23}$$

where $\sigma^{(n)2}_{ab}$ denotes the (annualised) multi-period n-year covariance between assets and benchmark portfolio

$$\begin{aligned} w^{(n)}_{\text{Active}} &= \arg\max E(r^{(n)}_{\text{Active},\,t+n}) - \tfrac{1}{2}(1-\gamma)\operatorname{var}(r^{(n)}_{\text{Active},\,t+n}) \\ &= \frac{1}{\gamma}\left[\left(1-\frac{1}{\gamma}\right)\Omega^{(n)}_{aa} + \frac{1}{\gamma}\Omega_{aa}\right]^{-1}[\mu^{(n)} + \tfrac{1}{2}\sigma^2_a - (1-\gamma)\sigma^{(n)}_{ab}] \end{aligned} \tag{15.24}$$

Again, the optimal solution can be decomposed into speculative demand and hedging demand. This case is almost identical to the previous case. Very risk-averse investors would hedge out benchmark-related risks, ie, for large γ we get

$$w^{(n)}_{\text{Active, hedging}} = [\Omega^{(n)}_{aa}]^{-1}\sigma^{(n)}_{ab} \tag{15.25}$$

If we interpret the benchmark portfolio as pension liabilities, Equation 15.25 would require investment in closely correlated (liability matching) assets.

15.4.3 Investor with non-tradable wealth

So far we have looked at an asset-only problem and an asset-liability problem. Finally, we introduce non-tradable real wealth. Assume that an investor not only can allocate assets in their financial wealth but also has non-financial wealth that is not tradable. We can think of this as human capital (slavery is abolished although divorce laws constantly contradict that statement) in individual portfolio choice or underground oil wealth in sovereign wealth fund asset allocation.

We assume a resource-based sovereign wealth fund. Sovereign wealth consists of both financial assets (the sovereign wealth fund) and non-tradable, non-financial assets, ie, underground oil. Let θ denote the fraction of financial wealth to total wealth. By now we can simply write down risk, return and objective function as

$$E(r^{(n)}_{\text{sw},t+n}) = \theta w_t^{(n)T}(\mu^{(n)} + \tfrac{1}{2}\sigma_a^2) - \tfrac{1}{2}\theta w_t^{(n)T}\Omega_{aa}\theta w_t^{(n)} \tag{15.26}$$

$$\operatorname{var}(r^{(n)}_{\text{sw},t+n}) = \theta^2 w_t^{(n)T}\Omega^{(n)}_{aa} w_t + 2\theta(1-\theta)w_t^{(n)T}\Omega^{(n)}_{\text{a,o}} + 2\theta w_t^{(n)T}\Omega^{(n)}_{ac} \tag{15.27}$$

$$w^{(n)}_{\text{sw}} = \arg\max E(r^{(n)}_{\text{sw},t+n}) - \tfrac{1}{2}(1-\gamma)\operatorname{var}(r^{(n)}_{\text{sw},t+n}) \tag{15.28}$$

Here $r^{(n)}_{\text{ntw},t+n}$ denotes the log of sovereign wealth return, while $\Omega^{(n)}_{\text{a,o}}$ represents the vector of covariances between financial assets and changes in resource-based (oil) wealth.

Taking first derivatives from Equation 15.28 we get

$$\begin{aligned}\theta(\mu^{(n)} + \tfrac{1}{2}\sigma_a^2) - \theta^2\Omega_{aa}w_t^{(n)} + \tfrac{1}{2}(1-\gamma)\theta^2 2\Omega^{(n)}_{aa}w_t^{(n)} \\ + \tfrac{1}{2}(1-\gamma)2\theta(1-\theta)\Omega^{(n)}_{\text{ao}} + \tfrac{1}{2}(1-\gamma)2\theta\Omega^{(n)}_{ac} = 0\end{aligned} \tag{15.29}$$

We divide by θ, collect all terms involving $w_t^{(n)}$ on the left-hand side and use the identity that $(1-\gamma) = -(\gamma-1)$ to finally arrive at

$$\begin{aligned} w_{\mathrm{ntw}}^{(n)} &= \frac{1}{\gamma}\left(\frac{1}{\gamma}\Omega_{aa} + \frac{\gamma-1}{\gamma}\Omega_{aa}^{(n)}\right)^{-1} \\ &\qquad \times \left[\frac{1}{\theta}(\mu^{(n)} + \tfrac{1}{2}\sigma_a^2) - (\gamma-1)\frac{1-\theta}{\theta}\Omega_{\mathrm{ao}}^{(n)} - \frac{1}{\theta}(\gamma-1)\Omega_{ac}^{(n)}\right] \end{aligned} \tag{15.30}$$

The optimal portfolio now contains three-fund separation, ie, speculative demand as well as hedging demand against changes in short rates (reinvestment risk) and changes in oil-related wealth

$$w_{\mathrm{sw,\,speculative}}^{(n)} = \left(\frac{1}{\theta}\right)\left(\frac{1}{\gamma}\right)\left(\frac{1}{\gamma}\Omega_{aa} + \frac{\gamma-1}{\gamma}\Omega_{aa}^{(n)}\right)^{-1}(\mu^{(n)} + \tfrac{1}{2}\sigma_a^2) \tag{15.31}$$

$$w_{\mathrm{sw,\,cash\text{-}hedge}}^{(n)} = -\left(\frac{1}{\theta}\right)\left(\frac{\gamma-1}{\gamma}\right)\left(\frac{1}{\gamma}\Omega_{aa} + \frac{\gamma-1}{\gamma}\Omega_{aa}^{(n)}\right)^{-1}\Omega_{ac}^{(n)} \tag{15.32}$$

$$w_{\mathrm{sw,\,oil\text{-}hedge}}^{(n)} = -\left(\frac{1-\theta}{\theta}\right)\left(\frac{\gamma-1}{\gamma}\right)\left(\frac{1}{\gamma}\Omega_{aa} + \frac{\gamma-1}{\gamma}\Omega_{aa}^{(n)}\right)^{-1}\Omega_{\mathrm{ao}}^{(n)} \tag{15.33}$$

For $\theta = 1$ we arrive at the optimal solution for an investor with financial wealth only. For $\gamma \to \infty$ we get

$$\begin{aligned} w_{\mathrm{sw,\,cash\text{-}hedge}}^{(n)} &= -\left(\frac{1}{\theta}\right)[\Omega_{aa}^{n}]^{-1}\Omega_{ac}^{(n)} \\ w_{\mathrm{sw,\,oil\text{-}hedge}}^{(n)} &= -\left(\frac{1-\theta}{\theta}\right)[\Omega_{aa}^{(n)}]^{-1}\Omega_{\mathrm{ao}}^{(n)} \end{aligned}$$

Also it is interesting to note that, for $\gamma = 1$, ie, a log investor, we arrive at the familiar solution that there is no hedging demand and the optimal solution degenerates to the one-period myopic speculative demand

$$w_{\mathrm{sw,\,speculative}}^{(n)} = \frac{1}{\theta}\Omega_{aa}^{-1}(\mu^{(n)} + \tfrac{1}{2}\sigma_a^2)$$

15.4.4 An extended example

Let us further extend our sovereign wealth fund (SWF) example. We proxy the investment universe for an SWF contained in x_t by the CRSP value-weighted stock index, real estate investments (FTSE NAREIT US Real Estate Index), government bonds (Lehman Long US Treasury Bonds Index), one-month Treasury bills rate and real bonds (TIPS). For changes in oil wealth, we use Brent current month

Table 15.3 Summary statistics

	Symbol	Mean (μ)	Volatility (σ)	Sharpe	Min	Max	Skew	Kurtosis
Assets								
Oil	ro_t	−0.37	19.23	−0.04	−0.89	0.82	−0.27	6.33
TIPS	rr_t	0.42	0.60	1.40	−0.02	0.02	−0.94	2.29
T-bill	rc_t	1.29	0.59		0.00	0.03	0.32	0.12
Long bonds	rb_t	1.33	4.97	0.54	−0.09	0.17	0.51	0.37
Equities	re_t	1.69	8.14	0.42	−0.28	0.18	−0.79	1.46
Real estate	ri_t	1.42	7.03	0.40	−0.17	0.19	0.00	0.05
State variables								
Nominal	yn_t	7.15	2.56		0.04	0.15	0.96	0.36
Credit spread	cs_t	1.96	0.50		0.01	0.04	1.04	0.67
Time spread	ts_t	1.75	1.25		−0.01	0.05	−0.05	−0.62
Dividend yield	dy_t	−3.66	34.24		−4.55	−2.97	−0.21	−0.58

We report descriptive statistics for all endogenous (assets and state) variables in our VAR. Only the Sharpe ratios have been annualised. The Sharpe value is calculated using $\sqrt{4}(\mu - rc)/\sigma$.

FOB. Given very high correlations between Brent, West Texas Intermediate and Arab Light, the choice of variable does not matter for our purposes. Economic state variables, s_t, are given by dividend yield, credit spread (the difference between Baa Corporate Bond Yield and the 10-year Treasury constant maturity rate), term spread (the difference between the 10-year Treasury constant maturity rate and 1-month T-bill rate) and nominal yield (10-year Treasury constant maturity rate).

Our analysis is based on quarterly returns from 1973 Q1 through 2007 Q4. Table 15.3 summarises the data and our notation. It is noteworthy that oil returns are statistically no different from cash rates, ie, excess returns are not statistically different from zero. This supports the Hotelling–Solow rule under perfectly integrated capital markets. Natural resource prices should grow at the world interest rate such that countries are indifferent between depletion (earning the interest rate) and keeping oil underground (earning price changes). Our state variables are common in the academic literature and motivated by time variation in investment opportunities and the investment universe will span most of the investment opportunity set for an SWF.

We start with calibrating our data generating process to the data in Table 15.3. Coefficient estimates and R^2 values are given in Table 15.5,

Table 15.4 Results from first-order VAR: residual covariance matrix

	rr_t	rc_t	rb_t	re_t	ro_t	ri_t	yn_t	cs_t	ts_t	dy_t
rr_t	**0.47**									
rc_t	0.05	**0.12**								
rb_t	−0.45	−0.08	**4.28**							
re_t	−0.28	−0.07	0.10	**7.11**						
ro_t	0.54	0.09	−0.22	−0.31	**17.84**					
ri_t	−0.40	−0.16	0.26	0.55	−0.28	**6.65**				
yn_t	0.20	0.08	−0.28	−0.15	−0.08	−0.17	**0.25**			
cs_t	−0.01	−0.19	0.02	−0.01	−0.01	0.01	−0.54	**0.18**		
ts_t	0.09	−0.33	−0.24	0.11	−0.21	0.02	0.47	−0.19	**0.43**	
dy_t	0.05	−0.08	0.04	−0.27	−0.02	−0.31	0.18	−0.21	0.03	**13.85**

We report the error covariance matrix, $\Omega_{\varepsilon\varepsilon}$. The main diagonal contains (quarterly) volatility while off-diagonal entries represent correlations.

while Table 15.4 provides estimates for Ω, with the main diagonal representing quarterly volatility. Unexplained quarterly real estate volatility amounts to 6.65%, which is only marginally smaller than unconditional volatility of 7.03%. This is hardly surprising given the R^2 value for the real estate equation is extremely low. Dividend yields have significant forecasting power for equities and bonds, while bonds can also be forecasted using last quarters nominal yields. Shocks on the dividend yields show negative contemporaneous correlation (−0.27) with equity returns.

At the same time rising dividend yields impact future equity returns positively (positive significant regression coefficient of 0.12). In summary, positive shocks to dividend yields (an increase) negatively impacts on current returns but positively impacts on next period returns. This creates mean reversion and is likely to reduce long-term risks. All state variables are very persistent as the high R^2 values and the large highly significant autoregressive coefficients indicate. Also, note that our system is stable, as all eigenvalues of $\hat{B}$ have modulus less than 1 (the largest is 0.94). Hence, the process is stable and therefore also stationary. The results also confirm recent work by Driesprong *et al* (2008), who found that lagged oil price changes predict future equity market returns. We find a t value of 2.55 on quarterly data, which is higher than their t values using weekly data.[6]

In this section we will focus exclusively on the hedging demand for an SWF as $w_{t,\text{spec}}^{(n)}$ is essentially the same for an asset-only investor,

Table 15.5 Results from first-order VAR: parameter estimates

	z_t									
	x_t						s_t			
	rr_t	rc_t	rb_t	re_t	ro_t	ri_t	yn_t	cs_t	ts_t	dy_t
a	0.04	−0.00	−0.41	0.57	0.62	−0.02	0.01	−0.01	0.05	−1.30
	0.01	0.00	0.11	0.18	0.46	0.17	0.01	0.00	0.01	0.36
	3.23	−0.61	−3.73	3.08	1.34	−0.12	2.13	−2.02	4.76	−3.64
rr_{t-1}	0.08	−0.06	2.28	−4.29	−5.77	−1.13	0.06	−0.02	0.14	−1.61
	0.13	0.03	1.14	1.90	4.76	1.77	0.07	0.05	0.12	3.70
	0.65	−1.97	2.00	−2.26	−1.21	−0.64	0.85	−0.37	1.24	−0.44
rc_{t-1}	−0.22	1.45	−8.42	−5.27	−7.36	−12.23	0.65	−0.22	−5.65	15.38
	0.44	0.11	3.97	6.59	16.53	6.16	0.24	0.17	0.40	12.83
	−0.51	13.51	−2.12	−0.80	−0.45	−1.99	2.76	−1.36	−14.04	1.20
rb_{t-1}	−0.02	−0.01	0.10	0.42	−1.25	0.13	−0.10	0.03	−0.03	−0.32
	0.01	0.00	0.11	0.18	0.44	0.16	0.01	0.00	0.01	0.34
	−1.51	−5.14	0.95	2.43	−2.86	0.78	−16.22	7.52	−2.88	−0.94
re_{t-1}	−0.01	0.00	0.12	0.05	−0.12	0.04	0.00	−0.01	0.00	−0.05
	0.01	0.00	0.07	0.11	0.29	0.11	0.00	0.00	0.01	0.22
	−1.71	0.10	1.75	0.45	−0.41	0.36	0.28	−3.41	0.71	−0.24
rb_{t-1}	−0.00	0.00	0.02	0.12	−0.09	0.04	0.00	−0.00	0.00	0.09
	0.00	0.00	0.03	0.05	0.12	0.05	0.00	0.00	0.00	0.09
	−0.71	0.75	0.52	2.55	−0.75	0.98	0.12	−0.19	0.54	1.01
ri_{t-1}	0.00	0.00	−0.16	−0.24	0.08	−0.01	0.01	−0.01	−0.01	0.04
	0.01	0.00	0.09	0.15	0.37	0.14	0.01	0.00	0.01	0.28
	0.50	0.67	−1.84	−1.63	0.23	−0.11	1.03	−2.16	−0.88	0.13

an asset liability investor or an asset-only investor with non-tradable wealth and hence not specific to an SWF. We ignore effects from γ and θ as they only affect leverage. The results are summarised in Figures 15.3 and 15.4. In both cases hedging demand is positive as long as correlation between investment and risk source is negative. In other words, assets that show negative correlation with oil or time-varying short rates will get positive weights. While correlation drives the sign of hedging demand, it is the relative term structure of risk (ie, whether annualised volatility grows or decays with the time horizon) that determines, all things being equal, the extent of hedging demand which essentially equals a "beta" estimate. For example, $[\Omega_{aa}^{(n)}]^{-1}\Omega_{\mathrm{ao}}^{(n)}$ describes a vector of asset betas relative to oil.

Long government bonds play the biggest role in both hedge portfolios as government bonds are both a recession hedge (they increase

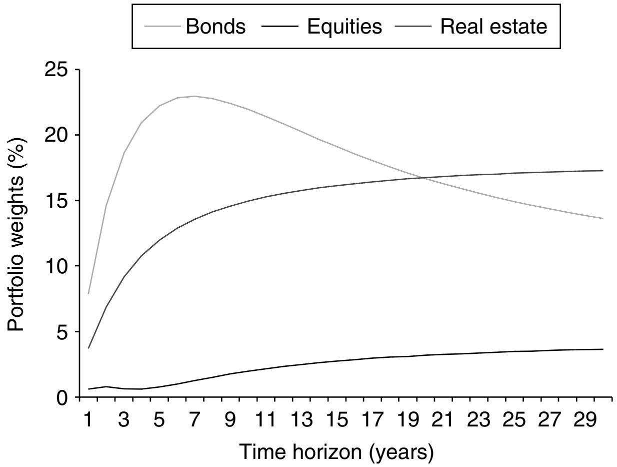
Table 15.5 (*continued*)

	z_t									
	x_t						s_t			
	rr_t	rc_t	rb_t	re_t	ro_t	ri_t	yn_t	cs_t	ts_t	dy_t
yn_{t-1}	−0.11	−0.11	3.33	−0.26	−0.96	2.71	0.80	0.08	1.17	−1.12
	0.11	0.03	0.95	1.58	3.96	1.48	0.06	0.04	0.10	3.07
	−1.05	−4.33	3.51	−0.17	−0.24	1.83	14.13	1.92	12.16	−0.37
cs_{t-1}	−0.02	−0.07	−1.12	−0.04	1.53	1.42	0.09	0.82	0.12	0.31
	0.11	0.03	0.99	1.65	4.13	1.54	0.06	0.04	0.10	3.21
	−0.18	−2.52	−1.13	−0.02	0.37	0.92	1.61	19.89	1.17	0.10
ts_{t-1}	0.06	0.18	−1.17	−0.80	−1.97	−1.91	0.14	−0.04	−0.60	6.74
	0.10	0.02	0.90	1.50	3.77	1.40	0.05	0.04	0.09	2.92
	0.63	7.33	−1.29	−0.53	−0.52	−1.36	2.53	−1.11	−6.59	2.31
dy_{t-1}	0.01	−0.00	−0.09	0.12	0.11	−0.00	0.00	−0.00	0.01	0.71
	0.00	0.00	0.02	0.04	0.10	0.04	0.00	0.00	0.00	0.08
	2.51	−0.98	−3.63	2.86	1.09	−0.07	2.28	−2.88	4.24	8.86
R^2	0.35	0.96	0.26	0.22	0.14	0.11	0.99	0.88	0.88	0.83
$\overline{R^2}$	0.28	0.96	0.18	0.14	0.04	0.01	0.99	0.86	0.86	0.81

We report the coefficients of a VAR(1) together with standard errors and t values. See Table 15.3 for a description of individual variable names. The last two rows contain the "R-square" as well as "adjusted R-square" for each individual regression. Coefficients significant at the 5% level are presented in bold.

in value when oil prices come down in a recession) as well as a hedge against deteriorating investment opportunities, ie, falling rates. While any empirical result is subject to estimation error, we observe that the sign of our hedging demand coincides with our economic prior. Government bonds are a recession hedge and tend to pay out in those states of the world were demand for commodities is low and hence prices will be depressed. While our data set ends at the end of 2007, we observe that this hedge would have worked extremely well in 2008, when all asset classes experienced record losses and oil fell from US$150 to below US$40 a barrel. Only Government bonds performed well, with the 10-year Treasury note up about 20%. Our analysis concludes that an SWF should hold a considerable amount of assets in long US government bonds in order to hedge both the negative effects of oil price shocks as well as deteriorating short rates. Empirically, we find confirmation in the fact that part of the growing current account surplus of commodity exporters has been invested in US Treasury bonds. However, with

Figure 15.3 Hedging time variation in short-term interest rates



We represent the hedge portfolio according to $w^{(n)}_{\text{Hedge, cash}} = -(1/\theta)[\Omega^{(n)}_{=aa}]^{-1}\Omega^{(n)}_{ac}$ for $n = 1, \ldots, 30$ and $\theta = \frac{1}{2}$.

SWF funds recording losses of about 40% in 2008, while government bonds yielded about 20% in the same period, they did not allocate nearly enough.

15.5 RECENT DEVELOPMENTS IN DYNAMIC PORTFOLIO CHOICE

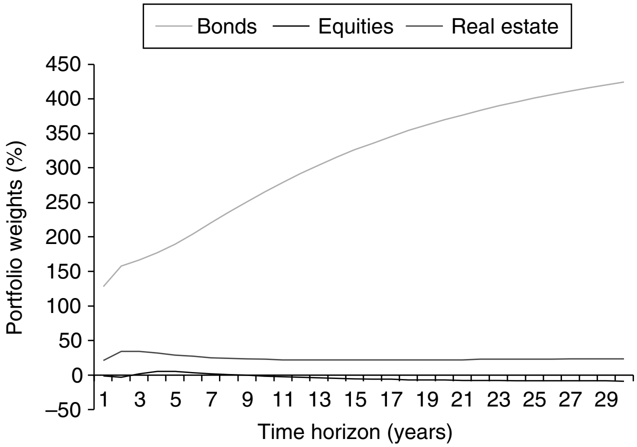
The above sections describe an explicit asset-allocation framework that ties investment decisions to both the conditional risk and return expectations at any point in time as well as the time horizon for investment decisions. However, this framework relies on return predictability which recently came under considerable attack. Return predictability is itself questioned on various grounds.

Ang and Bekaert (2006) ask the provocative question: "Is it really there"? They find that:

> At long horizons, excess return predictability by the dividend yield is not statistically significant, not robust across countries, and not robust across different sample periods. In this sense, the predictability that has been the focus of most recent finance research is simply not there.

All they find is short-term predictability when stock returns are conditioned on the short-term rate. This is confirmed by Goyal and

Figure 15.4 Hedging oil price shocks over time



We represent the hedge portfolio according to

$$w^{(n)}_{\text{Hedge, oil}} = -((1-\theta)/\theta)[\Omega^{(n)}_{aa}]^{-1}\Omega^{(n)}_{\text{ao}}$$

for $n = 1, \dots, 30$ and $\theta = \frac{1}{2}$.

Welch (2008), who find that existing models of return predictability fail to outperform a naive, uninformed iid model, both in- and out-of-sample.

The second line of criticism is parameter and model uncertainty. Even if predictability is found in-sample, how much can a real-time investor rely on the estimated parameters? Barberis (2000) and Brandt *et al* (2005) include parameter uncertainty and Bayesian learning in their model and find that parameter uncertainty weakens the positive link between time-horizon and equity allocations. It sometimes even reverses the outcome.

The third line of criticism is directly related to estimating the underlying VAR. Here the explanatory variables are highly persistent. Ferson *et al* (2003) argue that this leads to overstated degrees of predictability and potentially spurious regressions. After correcting for the persistence in right-hand variables (dividend yield, short-term rates, inflation, etc) predictability is usually much lower.

Finally, the estimated regression coefficients show little sign of stability (which is certainly related to the lack of out-of-sample performance). Results are highly subsample specific, as found by

Paye and Timmerman (2005), and suffer from small sample bias as discussed by Engsted and Pedersen (2008). This coincides with growing evidence for correlation breakdowns. If this is the case, we need to distinguish between economic regimes to better capture the prospects of asset classes in crash or boom periods. Regime switching models, as argued in Ang and Bekaert (2002) or Guidolin and Timmerman (2007), are being increasingly investigated. In these models, asset returns follow a more complicated process of multiple regimes with very different distributions of asset returns for each of them. Optimal investing under this class of models requires the formation of beliefs on which regime is active. By contrast, traditional asset-allocation techniques strongly assume that investors know with certainty which regime they are in.

[1] See Brandt *et al* (2005) for a review.

[2] See Samuelson (1963) and Rubinstein (2006) for a more detailed treatment of the above arguments.

[3] See Cochrane (2001, Chapter 20) for a review of earlier studies.

[4] I continue to strongly question this approach, as it defies first principles in corporate finance, but would refer the interested reader to Scherer (2005).

[5] See Lynch (2009) for R-code and the calculations described in this section.

[6] However, our regression coefficient shows the opposite sign but, given their regression left many statistically significant explanatory variables out, we should not be surprised if their results were biased.

REFERENCES

Ang, A., and G. Bekaert, 2002, "International Asset Allocation with Regime Shifts", *Review of Financial Studies* 15, pp. 1137–87.

Ang, A., and G. Bekaert, 2006, "Stock Return Predictability: Is It There?", NBER Working Paper 8207.

Barberis, N., 2000, "Investing for the Long Run when Returns Are Predictable", *Journal of Finance* 55, pp. 225–64.

Brandt, M., A. Goyal, P. Santa-Lara and J. Stroud, 2005, "A Simulation Approach to Dynamic Portfolio Choice with an Application to Learning about Return Predictability", *Review of Financial Studies* 18, pp. 831–73.

Campbell, J., and L. Viceira, 2002, *Strategic Asset Allocation: Portfolio Choice for Long-Term Investors* (Oxford University Press).

Cochrane, J., 2001, *Asset Pricing* (Princeton, NJ: Princeton University Press).

Driesprong, G., B. Jacobsen and B. Maat, 2008, "Striking Oil: Another Puzzle?", *Journal of Financial Economics* 89(2), pp. 307–27.

Engsted, T., and T. Q. Pederson, 2008, "Return Predictability and Inter-Temporal Asset Allocation: Evidence from a Bias Adjusted VAR Model", Working Paper, Aarhus University.

Ferson W. E., S. Sarkissian and T. T. Simin, 2003, "Spurious Regressions in Financial Economics?", *The Journal of Finance* 58(4), pp. 1393–413.

Goyal, A., and I. Welch, 2008, "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction", *Review of Financial Studies* 21(4), pp. 1455–1508.

Guidolin, M., and A. Timmerman, 2007, "Strategic Asset Allocation and Consumption under Multivariate Regime Switching", *Journal of Economic Dynamics and Control* 31, pp. 3503–44.

Lettau, M., and S. C. Ludvigson, 2001, "Consumption, Aggregate Wealth and Expected Stock Returns", *The Journal of Finance* 56, pp. 815–49.

Lynch, S. 2009, *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists* (New York: Springer).

Menzley, L., T. Santos and P. Veronesi, 2004, "Understanding Predictability", *Journal of Political Economy* 112, pp. 1–47.

Merton, R., 1971, "An Intertemporal Capital Asset Pricing Model", *Econometrica* 41(5), pp. 867–89.

Paye, B. S., and A. Timmerman, 2005, "Instability of Return Prediction Models", *Journal of Empirical Finance* 13(3), pp. 274–315.

Rubinstein, M., 2006, *A History of the Theory of Investments* (Chichester: John Wiley & Sons).

Samuelson, P. A., 1963, "Risk and Uncertainty: A Fallacy of Large Numbers", *Scientia* 98, pp. 1–6.

Scherer, B., 2005, *Liability Hedging and Portfolio Choice* (London: Risk).

16

Risk Management for Asset-Management Companies

"For an asset manager the greatest risk is operational risk" (Hull 2007, p. 372). In 2008, however, asset-management companies came under severe profitability pressure from market rather than operational risks. What has been seen as an annuity stream that was thought to expose firms to little or no earnings risk materialised as directional stock market exposure combined with high operational leverage (high ratio of fixed to variable costs). While operational leverage led to what has been praised as a scalable business (low costs of taking on additional business), in good times it was always clear that this would lead to massive losses in bad times. In short, asset managers partially share a client's benchmark risks. As client benchmarks went down, so did asset-based fees (percentage fee applied on average assets under management (AUM) within a year), which still represent the bulk of fee agreements in the asset-management industry.

At the same time, operational leverage increased the downturn in profits. A small example should make the mechanics clear. Suppose an asset manager with US$100 billion AUM, 50bp fees and 35bp total costs and an operational leverage of 90% (31.5bp in fixed costs). At the outset, the expected profits for the year are US$150 million. For a benchmark volatility of $\sigma = 30\%$ asset-based fees will be exposed to a 17% variability (see the next section for the plausibility and calculation of these numbers). In other words, asset-management revenue will be down by about 35% in a 2σ event. Should average AUM fall by 35%, all profits are wiped out and the company is left with a loss of US$12.75 million.[1] A 35% reduction in revenues led to a reduction in profits of more than 100%. Operational leverage leads to a reduction in profits that is many times larger than the reduction in revenues (fees). Note that this number assumes zero

Table 16.1 Volatility of average stock prices

σ (%)	Approx. 16.1 (%)	Approx. 16.2 (%)	Bootstrapping (%)
50	31.1	28.9	30.5
40	24.2	23.1	24.1
30	17.8	17.3	17.9
20	11.7	11.5	11.9
10	5.8	5.8	5.9

redemptions and zero shifts from high-fee equity products to low-fee fixed-income funds.

Given the size of this result it is surprising that in 2008 asset managers did not undertake any effort to reduce a source of risk that was outside their control. In fact, since then, year after year careful business plans have been drafted with detailed planning on new flows and revenues coming from existing and new clients, distribution channels, etc, while markets continued to make a complete mockery out of these exercises. Even a $\pm 1\sigma$ event on market returns leads to windfall gains or losses that are outside the control of an asset-management firm. While we could make the point that asset managers have a competitive advantage in assessing and taking stock market risks,[2] this point would also have required them to actively manage these risks over time. They did not.

The chapter is organised as follows. The next section introduces some closed-form solutions to approximate the volatility of asset-management fees as well as providing examples for the closeness of these approximations. We then test our conjecture that asset management is not an annuity business for T. Rowe Price. Next we try to explain the failure of actively hedging fees-at-risk (FaR) as a combination of misapplications of financial theory as well as corporate governance issues. We then review the case for hedging and make some attempt to lay out what and how to hedge.[3] Two appendices contain technical material.

16.1 FEES AT RISK

In order to assess the potential impact of market exposure in an asset manager's profit-and-loss (P&L) account, we need to find an expression for the volatility of asset-management fees. Given asset-based fees are calculated as a percentage of the average AUM over a time period, the calculations become slightly more involved than the usual value-at-risk calculations. Note that asset-based fees contain both benchmark (beta) and non-benchmark (alpha) exposure. We will always focus (unless stated otherwise) on the client-imposed benchmark part of asset-based fess. Implicitly, we assume that the alpha part of asset-based fees is negligible, which will be true for most mandates. Using standard results from derivatives pricing, we can approximate the volatility of annual asset-management revenues by[4]

$$\sigma(\tilde{f}_{t+n}) \approx \theta \cdot A_t \sqrt{\left(\frac{2\mathrm{e}^{\sigma^2} - 2(1+\sigma^2)}{\sigma^4}\right)\left(\frac{2\mathrm{e}^{\sigma^2} - 2(1+\sigma^2)}{\sigma^4} - 1\right)} \tag{16.1}$$

where A_t denotes AUM at time t, θ reflects percentage fees and σ stands for the volatility of asset returns underlying the calculation of average AUM. For those that prefer simpler formulas, we could also use

$$\sigma(\tilde{f}_{t+n}) \approx \theta \cdot A_t \frac{\sigma}{\sqrt{3}} \tag{16.2}$$

which will provide very similar results. Instead of using the arithmetic average, it is based on the geometric average (Nelken 1996). An example will illustrate both the quality of the approximations and the extent of FaR in asset-management companies. Table 16.1 calculates the volatility of average asset prices under approximations in Equations 16.1 and 16.2 as well as the "true" bootstrapped volatility.

Both approximations work remarkably well compared with the "true" value from bootstrapping. The simpler approximation in Equation 16.2 seems, however, to persistently underestimate the volatility of asset-management fees. For smaller volatilities, the approximations become increasingly better. Fee volatility for a European equity mandate with 30% volatility (not unusual in 2008) is around 18%: too large to be left unmanaged. The according P&L volatility can easily translate into organisational instability due to key staff defections. Clients prefer stable and profitable asset-management firms that can follow long-term investment strategies due to their ability to maintain key staff and to invest into the required infrastructure and IT. Large swings in asset-management profitability are often accompanied by asset outflows, poor client perception, consultant downgrades, questions raised about independence and short-term cost reductions that often come at client expense. The 2008 financial year set a perfect example for these frictional costs of P&L volatility. We will provide a deeper understanding of these forces in the following sections.

16.2 CURRENT ORTHODOXY: WHY DID ASSET MANAGERS FAIL TO HEDGE THEIR P&L?

Asset-management companies have collectively failed to protect their revenues against a downturn in markets. Rather than being an annuity business, asset-management firms shared their client's benchmark risks and while their clients would actively manage that exposure, asset managers have not. What zeitgeist caused this risk management failure?

The first argument against hedging runs as follows. Holding financial assets requires an *ex ante* positive risk premium (markets go up on average). Hedging asset-based fees (that otherwise also would go up on average over time with markets) will create a long-run opportunity loss to shareholders. Why give up on this long-run windfall gain that upwards-drifting markets would provide? How do we address this concern? Let us review the basics of corporate finance. What projects should firms undertake? It is accepted knowledge in corporate finance that companies need not worry about their shareholders. Companies (or, better still, their management) need to only worry about engaging in positive net present value (NPV) projects. Does the positive expected fee growth from market-return-driven AUM growth represent a positive NPV project? The clear answer is no. This applies no matter how large the expected growth rate of assets might look. The reason for this (at first glance) unintuitive proposition is that asset-based fees are the most simple form of a derivative contract (with AUM as underlying). We know that the value of a derivative contract is independent of the real-world growth rate of the underlying. Thus, the growth rate does not matter. This is just rephrasing the fact that capital market investments provide zero NPV, or as Ross (2005, p. 71) put it: "since the fee is contingent on asset value as a contingent claim, its current value is independent of expected rates of returns."

The second argument is slightly more sophisticated, as it relates to another cornerstone of financial economics: the famous Modigliani–Miller (MM) argument for hedging irrelevance. In frictionless markets, ie, in the absence of (frictional) bankruptcy, costs or taxes hedging would be irrelevant. So why hedge asset-management fees (or oil price risk for airlines, or dollar risk for German sports car manufacturers)? Could shareholders of an asset-management company not just simply undo the implicit beta exposure that comes with asset-based fees in their private portfolios? The trouble with the MM argument is that it rests on unrealistic assumptions. In reality, markets are not without frictions. Hedging, for example, preserves costly liquidity or reduces frictional bankruptcy costs. These are particularly high in the banking industry. The financial year 2008 undoubtedly showed this, with an unprecedented number of bank runs. In other words, the MM argument is itself particularly irrelevant in the banking industry.[5] Reducing volatility in asset-management earnings reduces not only frictional bankruptcy costs but also the need to tap capital markets when they are the least willing to provide external financing.

The final argument that the author has been confronted with is what has been called intellectual risk. In other words, a hedging policy might be intellectually sound but, as soon as the hedge itself makes losses, corporate memory seems to fade and nobody sees the offsetting gains of the hedged position any more. Thus, the hedging policy might hedge corporate financial risk but not individual career risk. Just ask those airline treasurers that hedged their fuel costs at an oil price of US$140 per barrel. This argument is particularly relevant at the time of writing in 2009, when markets are perceived to be at their bottom. While the author has sympathy with the risk manager in an asset-management organisation with intellectual risks, this argument seems more rooted in positive (why things are done) rather than normative (how should things be done) theory. Intellectual risk seems more of a corporate governance issue than a serious argument against hedging FaR.

Finally, it must be said that it is certainly tempting for a chief investment officer (CIO) to be paid a large bonus on the basis of beta exposure and that the incentive to remove this exposure might be in the shareholders' interest but not in the CIO's interest.[6] Managers that get paid with options (nonlinear exposure to earnings variability) on the underlying business will see an increase in the value of their executive options if they fail to hedge the P&L against market risks. Hence, senior management often has good reason not to hedge FaR, even if (as we will argue in the next section) this will on average reduce shareholder value. Again, this is a corporate governance issue. The opposite is true for private partnerships (manager and owner coincide), where managers have a linear exposure to earnings variability. Hedging here would reduce income volatility and hence increase manager (and owner) utility.

16.3 NEW ORTHODOXY: WHY SHOULD ASSET MANAGERS HEDGE THEIR P&L?

It is well known that hedging is shareholder positive if it creates a positive NPV project.[7] Building on the previous section we argue that, while asset-based fees are zero NPV projects (not hedging them will create no negative NPV per se), they still create P&L risks that, in a world with capital market frictions and taxes, are costly. First, not hedging asset-based fees will remove positive NPV projects (investing in new products, people, IT platforms, etc) by the necessity to hold cash against these risks in order to maintain a target rating. Real projects like the new distribution office in Madrid, the new product development team or the marketing campaign for a successful product get crowded out. If, alternatively, no cash is held as risk capital, not hedging P&L risk will increase the expected value of frictional bankruptcy costs and simultaneously limit the ability to leverage (and hence the ability to reap a tax shield or to use operational leverage). Unhedged swings in fee income will also increase the value of the tax option the government holds against the asset-management company. Taxes have to be paid if profits are made, but with limited carry forwards and backwards no equal amount is received if losses are made. The larger these swings are, the higher the value of this option. This argument obviously depends on whether the tax option is at-the-money.

Hedging P&L risk due to capital market movements should also allow an improved observability of effort in the principal–agent relationship between firm management and shareholder. In other words, a CEO might have worked very hard and made all the right decisions but, despite all of this, end-of-year P&L is still negative due to falling equity markets. How can the CEO convince the board that they should be paid incentive compensation when the board tells them that they are lucky to get another year to prove their leadership? Not only might management compensation driven by windfall gains and losses attract the less skilled (who would take risks they cannot control if they can avoid it?), but it might also discourage effort, as it is unlikely to be reliably observed by shareholders.

While we could think of many other channels through which risk management will increase value, 2008 taught the industry an unforgettable message. Losses on asset-based fees are largest in a severe equity downmarkets, where almost all asset classes will fall in value, and clients will massively redeem assets to de-risk or raise much needed cash. And it is precisely in these states of the world that bank funding will also dry up. In other words, a high correlation between revenue risks and funding risks clearly calls for hedging FaR. After all, hedging protects costly liquidity, as losses exhaust internal capital that is much needed (preferred) to finance new projects (pecking order theory).

16.4 CASE STUDY: REVENUE SENSITIVITIES OF T. ROWE PRICE

If asset-management companies fail to hedge their risks, they will show a considerable beta exposure to risky assets, ie, asset management will cease to be the annuity business it has been perceived as. This conjecture has not been backed up by empirical evidence. In this section we therefore want to estimate the relationship between asset-management revenues and stock market returns. We decide to use quintile regressions (QRs) as well as the ordinary-least-squares (OLS) method to obtain a summary for this relationship across quintiles. In particular, we use this approach because we are interested in whether data in the tails of the conditional distribution exhibit a different sensitivity to changes in market returns from data around the central location. This would be very useful to know when calculating "FaR".[8]

We select T. Rowe Price as the only publicly listed pure asset-management company with long enough history. For example, Schroders is publicly listed, but it became an asset-management company only in 2001; State Street and Northern Trust also run banking and custody businesses and Janus has only few years of data available. Our data ranges from 2007 to 1988. As a proxy for fee

Table 16.2 OLS and QR for annual revenue growth versus two measures of annual market performance

Market return	Sensitivities	T. Rowe Price	
$r = \frac{1}{S_t}\left(\frac{1}{250}\sum_{i=1}^{250} S_{t+i}\right) - 1$	β_{OLS}	0.94	(3.81)
	$\beta_{0.3}$	1.37	(3.28)
	$\beta_{0.5}$	0.71	(1.65)
	$\beta_{0.7}$	0.45	(2.13)
	$\beta_q \overset{?}{=} \bar{\beta}$	3.03	0.06
$r = \frac{S_{t+250}}{S_t} - 1$	β_{OLS}	0.42	(2.79)
	$\beta_{0.3}$	0.45	(1.48)
	$\beta_{0.5}$	0.37	(1.68)
	$\beta_{0.7}$	0.25	(1.7)
	$\beta_q \overset{?}{=} \bar{\beta}$	0.3	0.74

Regression of annual revenue growth versus two definitions of market returns for 1988–2007. *Source:* FactSet/Datastream.

income we take annual revenue growth from Compustat/FactSet.[9] As a measure of market returns, r, we use both annual returns (for the MSCI World Index in US dollars), ie, $r = (S_{t+250}/S_t) - 1$ as well as annual percentage changes in average price relative to prices at the start of a year, ie

$$r = \frac{1}{S_t}\left(\frac{1}{250}\sum_{i=1}^{250} S_{t+i}\right) - 1$$

We conjecture the later better represents the nature of asset-based fees. Our results are documented in Table 16.2.

The revenue sensitivity in an OLS regression[10] for our standard definition of market returns is 0.42 with a significant t value of 2.79. This is a clear indication that asset management is not an “annuity business” but rather comes with substantial cyclical equity market exposure in its revenues. This is not to be confused with the stock market beta of T. Rowe Price. For example, the average annual beta of T. Rowe Price over the last 20 years has been about 1.58: hardly an annuity business but also driven by leverage. The above measure aims at the impact of markets on revenues directly.

However, the sceptical reader might reply that 0.42 looks like a “low number”. Let us investigate this claim. We can proof under fairly standard assumptions (see Appendix C on page 468) that fee sensitivities β_{fee} in an OLS regression will converge against half the asset class beta, β, as averaging becomes continuous ($n \to \infty$), ie

$$\beta_{\text{rev}} = \frac{(1 + \frac{1}{2}(n-1))\beta\sigma_{\text{m}}^2 n}{\sigma_{\text{m}}^2 n^2} = \frac{\beta(1 + \frac{1}{2}(n-1))}{n} \overset{n\to\infty}{=} \tfrac{1}{2}\beta \qquad (16.3)$$

In other words, the averaging process for fee income hides some market exposure. If we use returns on the average price as market return, the sensitivity roughly doubles (to 0.94) and t values increase. The low beta results from the averaging process and is not an indication of low market exposure. We also find that betas in the more extreme quintiles ($q = 0.3, 0.5$) provide more significant t values. The zero hypothesis that all quintile betas are equal can be rejected in this case with a p value of 6%.[11]

Ideally, we would prefer a zero equity beta in our revenue stream. Remember that beta exposure carries no implicit reward (zero NPV), but it comes at deadweight volatility costs (described in the previous section) ie, eventually a negative NPV. Some readers might now ask themselves: don't all businesses have some beta sensitivity in their revenue stream? Demand will rise if the economy is doing well and vice versa. True, but, firstly, some of that beta is part of the firm's core risk-taking activities. Extend your sports car production in a boom; reduce your lorry production before demand for cars (recession) falters. Secondly, this relation is fuzzy and indirect for most business, while the way asset-management fees are calculated creates a direct one-to-one relationship between market risk and revenue risk that is not desirable from a shareholder's perspective.

16.5 WHAT TO HEDGE?

So far, we have talked about why hedging P&L risk is NPV positive for an asset manager, but we have been very narrow (we focused on the beta part of asset-based fees, assuming no redemptions or client-specific risks) about exactly what we should hedge. This section tries to provide more context. Let us divide risks into production risks and business risks to get some further insight.

Production risks are risks that come as a by-product of managing client money. Both asset-based and performance-based fees fall into this category. Production risks come both as alpha (outperformance versus a risk-adjusted benchmark) and beta (economic factor exposure) risks. Taking (owning) alpha risks is one of the core competencies of an asset-management firm. The asset manager is rewarded for its scarce and industry-specific skills. Acquiring these risks is essentially a positive NPV project, as risks are more than compensated by the expected profits. Creating alpha is equivalent to creating positive NPV, both for the firm and its clients. Beta risks, on the contrary, are incidental risks that asset managers usually do not take either because they think they lack the skills. Even if asset managers were to exhibit skills in market timing, the risk–return ratio is likely to be so low that it seems wiser to hedge these risks to free-up risk capacity for more core, ie, higher NPV-generating, risks.

While we argued very strongly in favour of hedging the beta part of asset-based fees, it should be clear that we should not hedge performance-based fees. We would simply destroy the option value provided by the client. Here we want volatility. Also note that we can, by definition, only hedge systematic beta risks and that idiosyncratic risks are an asset manager's core product, ie, the core production risk to take. Practically, it would also be (legally) difficult to hedge performance-based fees, as this would require the asset manager to hold positions (for their own P&L in a separate brokerage account) offsetting those that they implement in a fiduciary function for a given client. Note that option-pricing technology becomes a dangerous tool here, as replication is impossible.

While it is not common for banks to also hedge business risks (risks that are common to a business model but not directly related to production) it is clear that some of these business risks are correlated with capital market risks. Business risks in the asset-management industry are mainly related to systematic outflows affecting whole product lines. In a severe equity downmarket, retail investors will shift their asset allocation out of fee-intensive equity funds into money market funds or government guaranteed deposits, while institutional clients like insurance companies might also reduce their risk exposures due to their own (now binding) regulatory constraints. We could think of various ways to hedge client redemption risk. Redemptions in the retail sector are usually correlated with asset performance and we might want to hedge against extremely bad market scenarios under which redemptions are likely to be triggered. For institutional clients, redemptions might additionally be motivated by the client's own financial distress, ie, the client might need to raise cash or de-risk its asset allocation. All instruments related to financial distress could be used. For example, an asset-management company that is exposed to a weakly rated insurance company might want to buy puts or credit default swaps on the insurer to hedge parts of its fee income. A fund of the hedge fund provider might want to buy puts on the hedge fund, replicating clones in order to isolate itself partially from client redemptions or FaR.

16.6 HOW TO HEDGE?

The easiest way to hedge asset-based fees is not to offer them. This follows the idea of duality in risk management. We can either root out the cause (variability in markets) or the effect (offer fixed fees). Fixed fees have been discussed in institutional asset-management for a while, but they are perceived to suffer from the obvious to renegotiate in a world with positive inflation (although they could be of course indexed). However the author anticipates an increase in these fee arrangements.

How do we practically implement a hedge programme aimed at insulating an asset manager's P&L from market-induced variations of its average AUM for a given time period? The naive proposition would be to sell futures with one-year maturity on the underlying assets with a notional $\theta \cdot A_t$. If assets increase in value, the hedge (ignoring carry) creates a loss of $-\theta \cdot \Delta A$, while asset-management fees rise by an offsetting $+\theta \cdot \Delta A$. However, this hedge will not generally work due to the path dependency of fees based on average AUM. We can easily think of a situation where equity markets have been falling gradually over 11 months in the year but sharply recover at year-end, compensating for more than its previous losses. Here we would loose money on the hedge (the equity market is up on the year) and in revenues (average AUM are down too). What we need is a hedging policy instrument that moves with average prices rather than year-end prices. Fortunately, all we need is to trade (or replicate) a forward contract on the average stock price, ie, a contract that pays the average stock price at the end of the period (Asian forward, $F^{\text{Asian}}_{t,t+n}$). The price of such a contract to sell the average stock price (yet random) at a known price $\bar{S}$ is given by[12]

$$F^{\text{Asian}}_{t,t+n} = \mathrm{e}^{-r}\left(S_t \sum_{i=1}^{n} \frac{\mathrm{e}^{ri/250}}{n} - \bar{S}\right) \tag{16.4}$$

Note that this is independent of the distribution that the average price process might follow. We assume $n = 250$. Another way to think of Equation 16.4 and to create a position in an Asian forward is to buy a long position into an Asian call with strike $\bar{S}$ and one-year maturity and a short position in an Asian put with the same strike and maturity. A long call and short put provide a synthetic forward that comes at zero cost if the strike is at-the-money forward.[13]

How would this work? We assume current AUM of US$100 billion with fees of 50bp.[14] The current P&L to defend is €500 million or €507.56 million at 2009 forward prices (0.5% $\cdot$ 100 billion $\cdot$ $\sum_{i=1}^{n} \mathrm{e}^{ri/250}/n = 507.56$ million). Suppose the asset manager wants to isolate the P&L at current rates of 3% per annum against market exposure. Suppose the benchmark asset trades at an index level of 4,500, where each index point is valued at €2,500. Here $S_t = 4\,500 \cdot 2{,}500 = 1{,}125{,}000 =$ €1.125 milion. A forward contract to sell the average index level at $\bar{S} =$ €1.142 million is valued at zero at time t. We need to sell

$$\#\text{Asian forwards} = \frac{A_t\theta}{F_{t,t+n}^{\text{Asian}}} = 444.444$$

Now suppose the average index level drops at the end of the year to

$$\frac{1}{n}\sum_{i=1}^{n} S_{t+i} = 3{,}800$$

The payout from our short forward position is

$$444.444\left(\bar{S} - \frac{1}{n}\sum_{i=1}^{n} S_{t+i}\right) = €85.3533 \text{ million}$$

Together with asset-management fees of

$$\theta \cdot \frac{3{,}800}{4{,}500} \cdot €100 \text{ billion} = €422.22 \text{ million}$$

this amounts to total fees of €507.56 million, which is exactly what we sold the fee income for the coming year for. We have insulated the P&L. One last objection against hedging asset-based fees could creep in here. After all, fixing your revenues when your input costs are variable (inflation, competitive pressure to hire talent, etc) does not sound like a good idea. There are two answers to this. First, inflation expectations are incorporated in the forward curve for pricing $F_{t,t+n}^{\text{Asian}}$ (although we assumed for simplicity a flat curve above). Second, if input costs change any business needs to increase prises or increase efficiency. There is nothing special about asset management.

Table 16.3 Statistical properties of revenue growth data from 2000 to 2008

	Business	Mean	Volatility	Skew	Kurtosis
Schroders	AAM	0.02	35.75	−0.34	−0.02
American Capital	PE/A	57.81	16.15	0.69	0.25
Bank OF NY Mellon	AAM	13.87	24.16	1.10	2.71
Janus Capital	AAM	3.44	36.24	1.92	4.53
Julius Baer Hldg	AAM	21.63	43.49	0.95	0.55
Northern Trust	PAM	9.87	17.43	−0.39	−2.07
SEI Investments	AMI	15.78	17.94	1.20	1.49
State Street	AAM	13.11	15.86	−0.78	−0.92
Price (T. Rowe)	AAM	11.13	15.77	−0.97	−0.50

AAM, active asset management; AMI, asset management infrastructure; PAM, passive asset management; PE/A, private equity/alternatives.
Source: FactSet.

16.7 AN EMPIRICAL ANALYSIS OF THE ASSET-MANAGEMENT INDUSTRY

Our data set consists of 72 annual data points, ie, eight years of annual revenue growth data (from Compustat/FactSet) for nine publicly traded asset-management firms. We focus on data from 2000 to 2007 in order to avoid using the same return observation in our study that also motivated us to research this matter. All results become considerably stronger if 2008 is included. We focus on listed asset-management companies, as revenue data is not available for private partnerships or bank-owned asset-management companies. A summary of the data is given in Table 16.3.

The publicly listed asset-management companies in our sample cover active as well as passive investments, asset-management infrastructure and private equity alternatives. We sampled the companies in order to arrive at the maximum number of total observations. All results in this chapter remain the same for longer and narrower panels (fewer asset-management firms and with a longer overlapping time but fewer total data points).

First, observe that unconditional average revenue growth (per annum) varies considerably across asset-management firms, where Schroders (0% annual growth) and American Capital (58% annual growth) represent the extremes. The around-zero average revenue growth for Schroders in our data sample is hardly surprising given their substantial exposure to institutional (Asian) equity and balanced mandates with too little initial retail presence and too small fixed-income and alternatives franchises that found it difficult to stem the outflows from a very traditional equity business. The annual volatility of revenue growth data is on a similar level to global stock market volatility, which confirms our view that equity risk is a main driver of revenue risk. American Capital, on the other hand, was perfectly positioned as a private equity and alternatives house and hence created 58% annual revenue growth.

The MSCI World Index in US dollars is used as a proxy for stock market returns.[15] However, instead of calculating year-end returns $r = \mathrm{MSCI}_1/\mathrm{MSCI}_0 - 1$, we use

$$r_{\mathrm{av}} = \frac{1}{S_0}\left(\frac{1}{250}\sum_{i=1}^{250} S_{0+i/250}\right) - 1$$

ie, the annual percentage changes in average price relative to prices at the start of a year. We conjecture the latter to better represent the nature of asset-based fees. Asset-based fees are calculated as a fraction of average AUM for a given year. Appendix A (on page 465) proves that the regression beta of year-end stock market returns against revenue growth will lead to misleading stock market sensitivities. Under fairly standard assumptions fee sensitivities β_{rev} in an OLS regression will converge against half the asset class beta, β, as averaging becomes continuous ($n \to \infty$), ie

$$\beta_{\mathrm{rev}} = \frac{(1+\frac{1}{2}(n-1))\beta\sigma_{\mathrm{m}}^2 n}{\sigma_{\mathrm{m}}^2 n^2} = \frac{\beta(1+\frac{1}{2}(n-1))}{n} \overset{n\to\infty}{=} \tfrac{1}{2}\beta \tag{16.5}$$

where σ_{m}^2 denotes benchmark volatility. In other words, the averaging process for fee income hides some market exposure. If we use returns on the average price as market return, the sensitivity roughly doubles (to 0.94) and t values increase. The low beta results from the averaging process and is not an indication of low market exposure, as the averaging process effectively decouples asset returns and fees even though there is a deterministic one-to-one relationship. Without yet pooling information in our panel data set, we run $i = 1, \dots, n = 9$ separate regressions of market return versus revenue growth for $t = 2000, \dots, 2007$

$$\mathrm{rev}_{it} = \alpha_i + \beta_i \cdot \mathrm{rev}_{it} + \varepsilon_{it} \tag{16.6}$$

Table 16.4 Revenue beta

	Business	α		β		$\bar{R}^2$
Schroders	AAM	−2.04	(−0.21)	2.61	(2.32)**	47.38
American Capital	PE/A	58.30	(10.08)***	−0.61	(−0.94)	12.81
Bank of NY Mellon	AAM	12.72	(1.67)*	1.45	(1.68)	32.04
Janus Capital	AAM	3.27	(0.24)	0.22	(0.14)	0.33
Julius Baer Hldg	AAM	19.95	(1.35)	2.13	(1.27)	21.30
Northern Trust	PAM	9.20	(1.55)	0.84	(1.26)	20.89
SEI Investments	AMI	15.27	(2.36)**	0.65	(0.88)	11.51
State Street	AAM	12.28	(2.58)**	1.05	(1.95)*	38.88
Price (T. Rowe)	AAM	10.14	(2.57)**	1.27	(2.84)**	57.27

AAM, active asset management; AMI, asset management infrastructure; PAM, passive asset management; PE/A, private equity/alternatives. The table shows the regression coefficients for separate OLS regressions. Statistical significance at the 10%, 5% and 1% level is denoted by *, ** and ***, respectively.

where α_i is autonomous revenue growth arising from a better sales force, the ability to charge growing fees by upgrading clients from low- to high-fee products or constant product innovation. The slope coefficient, β_i, is a measure of revenue sensitivity to stock returns.

A large and significant coefficient is the opposite of what we would expect to see under the so-called "annuity business view" held by so many practitioners. It would also establish market risk (and not just operational risks) as an important factor. The size of β_i depends on the mix between asset-based and performance-based fees as well as the product and mandate mix of an asset manager.[16] The results for the separate regression are summarised in Table 16.4. Schroders has the largest revenue beta (2.6 with a t value of 2.32) in our sample, which should not surprise us given its position as an equity manager, while American Capital shows the least correlation with global equity markets, which also should not surprise us given its focus on non-directional alternative investments. The adjusted R-squared value gets as large 57% but also as low as 1% for Janus Capital Group, which had massive problems around 2004, when practices of market timing and late trading became known to investors. Janus therefore did not participate in the recovery following the great crash from 2000 to 2003 and hence looks "market neutral". This is a highly individual effect that motivates us to use panel data, as we hope that these effects average out across the units in our panel. Finally, there is a literature suggesting a nonlinear relation between fund flows and performance. Chevalier and Ellison (1997) and Sirri and Tufano (1998) provide evidence that investors chase outperforming funds but do not leave underperforming funds at the same rate. On a macro level we could find a similar phenomenon. Investors might increasingly buy equity products when markets go up. Revenues would then increase faster in rising equity markets as the AUM grow via market impact and fund flows. We therefore test for nonlinearity using the Ramsey Reset Test. However, we cannot reject the null hypothesis of linearity at the 10% level for all firms apart from T. Rowe Price.

The first model to be estimated is the classical pooling (stacking) model, where all parameter variability is removed, ie, $\alpha_1 = \cdots = \alpha_9 = \alpha$ as well as $\beta_1 = \cdots = \beta_9 = \beta$. This increases our degrees of freedom if it allows us to average across individual units in our panel

$$\text{rev}_{it} = \alpha + \beta \cdot \text{rev}_{it} + \varepsilon_{it} \tag{16.7}$$

All asset-management firms are treated as identical. They exhibit the same revenue sensitivity to market movements and the same flow of new client money at the same fee structures. This is hardly realistic and so we can next estimate the fixed-effects model

$$\text{rev}_{it} = \alpha_i + \beta \cdot \text{rev}_{it} + \varepsilon_{it} \tag{16.8}$$

where we allow for variation in intercepts (capturing hidden, ie, not included, variables like sales force strength, fee structure, etc) but not in slopes. This is essentially a dummy variable regression (pooled regression with nine dummy variables). As an alternative to Equation 16.8 we also estimate the random effect models. This allows us to let α_i vary randomly and hence become part of the error term

$$\text{rev}_{it} = \alpha + \alpha_i + \beta_i \cdot \text{rev}_{it} + \varepsilon_{it} \tag{16.9}$$

Here unknown explanatory variables enter the error term an as long as the error term remains uncorrelated to the explanatory variables; random effects are consistent while fixed effects are not. The final variation we look at is the so-called random coefficients model. It assumes that all regression coefficients (not only the intercept) are drawn from a distribution with constant mean, ie

$$\text{rev}_{it} = \alpha + \alpha_i + (\beta_i + e_i)\text{rev}_{it} + \varepsilon_{it} \tag{16.10}$$

The variance of ε_i determines the dispersion in individual regression slopes and hence determines the degree of shrinkage towards β. It is essentially a matrix weighted GLS estimator that will weight regression coefficients depending on the quality (residual variance) of each regression and usually applied when the data appear to be non-poolable.[17] We use it as a robustness check, ie, how would our conclusions look if the data could not be pooled?

Let us discuss our estimation results now. We start with the classical pooling model (the second column in Table 16.5). Average annual revenue growth has been 15% per year as an average for all nine asset-management companies, which is highly significant (1% level) and a reflection of the continuing stream of assets into retirement savings as well as the industry's product innovation into hedge funds, structured investment funds and other high-fee-earning products. The sensitivity of revenues to stock market movement is close to 1 and also statistically significant at the 1% level. There is significant market risk in asset-management fees. However, we need to check whether estimation of the classical pooling model is allowed. Are the assumptions of homogeneity across asset-management firms justified? This is can be tested by a Chow F-test (Baltagi 2005)

$$\begin{aligned} F &= \frac{\mathrm{RSS}_{\mathrm{CP}} - \mathrm{RSS}_{\mathrm{SR}}}{\mathrm{RSS}_{\mathrm{SR}}} \frac{\mathrm{df}_{\mathrm{SR}}}{\mathrm{df}_{\mathrm{CP}} - \mathrm{df}_{\mathrm{SR}}} \\ &= F(\mathrm{df}_{\mathrm{CP}} - \mathrm{df}_{\mathrm{SR}}, \mathrm{df}_{\mathrm{SR}}) \end{aligned} \tag{16.11}$$

where $\mathrm{RSS}_{\mathrm{CP}}$ are the residual sum of squares for the pooling model, $\mathrm{RSS}_{\mathrm{SR}}$ are the residual sum of squares from separate regressions (sum of each regressions residual risk times the number of observations) and $\mathrm{df}_{\mathrm{CP}}$, $\mathrm{df}_{\mathrm{SR}}$ are the degrees of freedom for each model. The test statistic is $F(16, 54) = 2.25019$ with a significance level 0.012, ie, the zero hypothesis of classical pooling is rejected.[18]

On the back of these results we next try a less extreme form of pooling, ie, the fixed-effect model represented by the third column in Table 16.5. We cannot provide a pooled estimate for the average (stock market independent) revenue growth (due to perfect collinearity of the nine dummy variables and a vector of 1s for the intercepts), but there seems to be considerable variation in the autonomous growth rate for each asset manager. This alone increases the adjusted R-squared value to 29.81% (up from 8.7% for the pooling model). Applying the Chow poolability test again

Table 16.5 Alternative panel data models

	Classical pooling (16.7)	**Fixed effects (individual) (16.8)**	**Random effects (individual) (16.9)**	**Random coefficients (16.10)**
α	15.45	—	15.45	15.96
	(4.5)***	—	(2.75)***	(2.49)**
β	1.06	1.06	1.10	0.99
	(2.79)***	(3.19)***	(3.38)***	(2.33)**
$\bar{R}^2$	8.7%	29.81%	35.7%	—

Statistical significance at the 10%, 5% and 1% levels are denoted by $*$, $**$ and $***$, respectively.

(this time we test separate regressions versus a fixed-effect model) we get $F(8, 54) = 0.91740$ with significance level 0.51, ie, the null (fixed-effect model) cannot be rejected. Testing against the classical pooling model, the appropriate F-test yields $F(10, 60) = 3.62$ and is highly significant, ie, we reject the null hypothesis of classical pooling. The next step in our specification search is to estimate a random-effects model that would allow us to infer the unknown average autonomous revenue growth. A random-effects model would be generally preferable (the estimator is efficient and consistent, while fixed-effects models are only consistent). This applies even more in our case, where we only deal with a subsample of the population and want to draw inferences on the whole population of asset-management companies. Note that both random- and fixed-effects models treat our main parameter of interest, the revenue beta, as fixed. The random-effects model provides estimates close to the fixed-effects model, again with highly significant parameter estimates, but it only ranks over the latter if parameter estimates are unbiased.

Bias arises from a correlation of explanatory variables and the error term. The error term picks up omitted variables. In our case this could be sales force quality, product innovation, etc. If those variables are correlated with stock market returns, the random-effects estimator is biased, while the fixed-effect estimator is not (omitted variables are caught in the intercept not the error term). We test for correlation between error term and explanatory variables using the Hausmann test (Verbeek 2004, p. 351). The value of the $\chi^2(1)$ test statistic is 0.04, ie, we confirm the existence of random effects. Finally, we remove the assumption of equal slope (revenue beta) across asset-management firms. We view this as a robustness test. Would our conclusions so far, "asset-management revenues are driven by market beta", still remain valid? The fifth column in Table 16.5 provides the answer. Even under the assumption of random coefficients, our beta estimate remains highly significant and close to unity.

Our results show that asset-management companies do not hedge market risks despite them being mechanically built into their revenue stream. This is surprising, given that market risks arise from the client's benchmarks choice (eg, US equities) and are therefore incidental to the asset manager's production process (alpha versus US equities) ie, the creation of value added relative to a client-chosen benchmark. However, the empirical results confirm our earlier theoretical observations.

We have shown that asset-management revenues carry substantial market risks. This challenges both the academic view in the risk management literature about the predominance of operative risks as well as the industry practice of not hedging market risks that are systematically build into the revenue generation process. We have also provided a list of typical arguments against hedging asset-management fees. In order for asset-management companies to return to an annuity model, these risks need to be managed more actively. Shareholders do not want to get exposure to market beta via holding asset-management companies; rather, they want to participate in their alpha generation and fund gathering expertise as financial intermediaries.

16.8 HOW TO HEDGE THE RISK OF FUND OUTFLOWS?

The asset-management-related risk management literature is focused on providing techniques to manage the investment risks for a given investment strategy.[19] However, in an industry where delegated asset management is the norm, the literature mainly deals with managing risks on behalf of the client, ie, managing client risks. Very little is known about managing an asset manager's own business risk. Business risks in the asset-management industry are mainly related to systematic outflows affecting whole product lines. To the extent that these redemption risks are positively correlated with financial distress (and hence higher funding costs as well as frictional bankruptcy costs), their effect is amplified.

Hedge fund outflows undermine both the economic profitability of the hedge fund manager and the value-added proposition of a given investment strategy. Outflows reduce the asset manager's asset base on which asset-based fees are calculated, ie, they have a first-order impact on the business, while redemptions on strategies with illiquid assets might trigger further losses via fire-sales and margin calls from prime brokers. Note that this also applies to funds with only modest leverage, as hedge fund clients represent the "short side" in a hedge fund balance sheet.

This chapter focuses on managing the risks of extreme outflows. It develops a static hedging model for a hedge fund that faces the risk of large redemptions if capital market conditions worsen. While practitioners have used out-of-the-money put options on hedge fund clones to hedge mainly performance-related tail risks, this chapter suggests using binary options on the volatility index VIX (as "quasi" Arrow–Debreu securities) to hedge out extreme outflows in times of market crisis. An extensive numerical example concludes.

How could a hedge fund manager possibly hedge against outflows? In an ideal world they would simply buy Arrow–Debreu securities, ie, securities that pay out one monetary unit if an outflow occurs and zero monetary units in all other states of the world. In reality, contracts of this form will not be written, as hedge fund outflows might be self-inflicted, for example, by bad performance or poor client service. As in a standard moral hazard problem, the hedge fund manager might simply have no incentive to try hard and perform if their outflows are insured. Alternatively, outflows could simply be engineered (derivative payments are induced by outflows today that are simply reversed tomorrow) to extract profits from a counterparty providing the hedge.

In order to implement a strategy against redemption risks a hedge fund manager needs to first and foremost look for an instrument that is sufficiently positively correlated with their outflows; yet it needs to be entirely outside their influence to avoid the repercussions of perceived moral hazard as discussed above. Two instruments come to the author's mind that are likely to hedge systemic risks outside the manager's control (without clouding their alpha generation or client service incentives). First, the hedge fund (or fund of fund) manager could buy puts on hedge fund clones that best resemble their strategy (mix). After all, bad performance will trigger or coincide with outflows.[20] This strategy was employed by some hedge funds at the height of the crisis in 2008.[21] Alternatively, the manager could buy out-of-the-money digital options that pay out one monetary unit if the VIX rises by more than a prespecified amount ΔVIX. The rationale for this is that strongly increasing risks often lead to both deteriorating hedge fund performance and increased client risk aversion. Both factors are likely to trigger hedge fund redemptions, are outside the a manager's control and represent an exogenous event (Anson 2002).

This chapter suggests buying a digital (deep out-of-the-money) call (D_t) on the VIX as a "quasi" Arrow–Debreu security. This shares the binary structure of Arrow–Debreu securities, as it will pay out one monetary unit ($D_{t+1} = 1$) if the VIX jumps by more than a critical amount Δ_{VIX} and pay out zero in all other scenarios ($D_t = 0$). We can formally write

$$D_{t+1} = \begin{cases} 1 & \text{for } \Delta\text{VIX} > \Delta_{\text{VIX}} \\ 0 & \text{for } \Delta\text{VIX} \leqslant \Delta_{\text{VIX}} \end{cases} \tag{16.12}$$

where $\Delta\text{VIX} = \text{VIX}_{t+1} - \text{VIX}_t$. Let us further define $v = \Pr(\Delta\text{VIX} > \Delta_{\text{VIX}})$, ie, the real-world probability that the VIX rises enough to trigger a payout on the digital option at $t + 1$. The price of such an option is given by

$$D_t = E(D_{t+1}) + \varepsilon = v + \varepsilon \tag{16.13}$$

where ε represents a risk premium beyond the expected payout. We refer to the digital call as a "quasi" Arrow–Debreu security as it sometimes pays even if we are not in the critical state and it sometimes does not pay even though we are in this critical state. In other words, Equation 16.12 represents the payout from a digital option that is imperfectly correlated to fund outflows, ie, while it pays out if the assets are withdrawn from the fund with probability

$$\theta^{+} = \Pr(\Delta\,\text{AUM} \leqslant \Delta_{\text{AUM}} \mid \Delta\text{VIX} \geqslant \Delta_{\text{VIX}}) \tag{16.14}$$

there are also instances where this option does not pay out despite the fact that assets are leaving the fund. This happens with probability

$$\theta^{-} = \Pr(\Delta\,\text{AUM} \leqslant \Delta_{\text{AUM}} \mid \Delta\text{VIX} < \Delta_{\text{VIX}}) \tag{16.15}$$

We can now use the above notation to fill a standard 2×2 contingency table in Table 16.6. For any given data set this table allows us to estimate all involved probabilities.

Table 16.6 Contingency table

	$\Delta\,\mathrm{AUM} \leqslant \Delta_{\mathrm{AUM}}$	$\Delta\,\mathrm{AUM} > \Delta_{\mathrm{AUM}}$	
$\Delta\mathrm{VIX} \geqslant \Delta_{\mathrm{VIX}}$	$\theta^+ v$	$(1-\theta^+)v$	v
$\Delta\mathrm{VIX} < \Delta_{\mathrm{VIX}}$	$\theta^-(1-v)$	$(1-\theta^-)(1-v)$	$1-v$
	$\theta^+ v + \theta^-(1-v)$	$1-\theta^+ v + \theta^-(1-v)$	1

Note that the entries in Table 16.6 depend on the choice of Δ_{VIX} and Δ_{AUM} that effectively determine the "borders" for each cell. In empirical applications the researcher often would like to minimise the "noise-to-signal ratio" given by

$$\frac{(1-\theta^+)v}{(1-\theta)} \bigg/ \frac{\theta^+ v}{\theta} = \frac{(1-\theta^+)}{\theta^+}\frac{1-\theta}{\theta} \tag{16.16}$$

with respect to Δ_{AUM} and Δ_{VIX}, which is equivalent to maximising θ^+, ie, the likelihood that the digital option pays out when needed. Note that Equation 16.16 used the fact that $\Pr(\Delta\,\mathrm{AUM} \leqslant \Delta_{\mathrm{AUM}}) = \theta = \theta^+ v + \theta^-(1-v)$. Without apology we write down the objective function for our hedge fund manager in standard form as

$$\begin{aligned} h^* &= \arg\max E(h(D_{t+1}-D_t) - 1_{t+1} f) \\ &\qquad - \tfrac{1}{2}\lambda E(h(D_{t+1}-D_t) - 1_{t+1} f)^2 \end{aligned} \tag{16.17}$$

where λ defines the managers risk aversion, while 1_{t+1} represents an indicator function that assumes the value 1 if $\Delta\,\mathrm{AUM} \leqslant \Delta_{\mathrm{AUM}}$.[22] When assets are redeemed at a faster rate $\Delta\,\mathrm{AUM}$ than a given threshold rate Δ_{AUM}, the hedge fund manager is assumed in the need to receive a lump sum payment of f in order to stay in business, either because this ensures confidence in their counterparties or because it avoids the need to liquidate less liquid positions. While Table 16.6 provides the probability of each event (combination of VIX changes and fund flows), Table 16.7 presents the payouts associated with buying h digital calls at a price $v + \varepsilon$.

With the knowledge of Tables 16.6 and 16.7, the solution to can easily be verified as

$$h^* = -\frac{\varepsilon}{\lambda(\varepsilon^2 + v - v^2)} + \frac{v\theta^+(1-(v+\varepsilon)) - (1-v)\theta^-(v+\varepsilon)}{\varepsilon^2 + v(1-v)} f \tag{16.18}$$

Table 16.7 Payout table

	Δ AUM $\leqslant \Delta_{\text{AUM}}$	Δ AUM $> \Delta_{\text{AUM}}$
$\Delta\text{VIX} \geqslant \Delta_{\text{VIX}}$	$h(1-(v+e))-1f$	$h(1-(v+e))-0f$
$\Delta\text{VIX} < \Delta_{\text{VIX}}$	$h(0-(v+e))-1f$	$h(0-(v+e))-0f$

As usual for this type of mean variance solution, h^* is given by a combination of speculative demand, ie, the first term in Equation 16.18 involving a risk aversion parameter and pure risk minimisation demand, ie, the second term in Equation 16.18. Note that $v\theta^+(1-(v+\varepsilon))$ denotes the probability weighted payout from a digital option (one minus the cost of the digital option) for the upper left-hand quadrant in Table 16.6, while $(1-v)\theta^-(v+\varepsilon)$ denotes the probability weighted loss from a hedge position that did not pay out but was costly to enter. For a positive h^* we require that the payout from getting it right (VIX jumps and assets are redeemed) must exceed the opportunity costs from getting it wrong (VIX does not jump even though assets are redeemed).[23]

Empirical data on hedge fund flows is not widely accessible, as data providers usually charge hefty sums for this data even if it is used purely for research purposes. In order to provide an accessible example, this chapter uses the data provided in Baquero and Verbeek (2009, p. 43, Table 1), which consists of 40 quarterly[24] growth rates of aggregate hedge fund flows (the hedge fund sample consists of 1,543 open-ended hedge funds from TASS with a minimum of six quarters of returns history) between 1994 Q4 and 2004 Q3. While the sample period is short, an extension into the years 2007, 2008 and 2009 is likely to improve our results, given the well-known large hedge fund redemptions that followed credit-crisis-induced VIX increases in 2008.

Data aggregated over many funds has the advantage of not being overly influenced by funds that experienced redemptions because of a general missing value proposition (ie, bad performance), legal problems, manager change, etc. All these effects diversify away on an aggregated level and are precisely what we are not interested in. We want to analyse hedges against systemic forces that affect hedge fund flows in general. The flow data used is hence a good starting point for a diversified provider of hedge fund strategies. Arguably, we could gain further insight by disaggregating the data into sub-strategies, as

Table 16.8 Hedge fund flows and VIX changes for $\Delta_{\text{VIX}} = 10$ and $\Delta_{\text{AUM}} = -3.514\%$

			Case			
Date	**Flow (%)**	**ΔVIX**	**1**	**2**	**3**	**4**
1994 Q4	−2.35	5.21	0	0	0	1
1995 Q1	−6.46	0.88	0	1	0	0
1995 Q2	−2.28	2.74	0	0	0	1
1995 Q3	−1.46	1.12	0	0	0	1
1995 Q4	−3.27	3.22	0	0	0	1
1996 Q1	0.50	1.82	0	0	0	1
1996 Q2	−1.07	6.44	0	0	0	1
1996 Q3	1.12	4.6	0	0	0	1
1996 Q4	2.60	2.48	0	0	0	1
1997 Q1	5.61	−0.61	0	0	0	1
1997 Q2	0.66	0.29	0	0	0	1
1997 Q3	4.71	3.08	0	0	0	1
1997 Q4	1.15	13.82	0	0	1	0
1998 Q1	2.95	4.03	0	0	0	1
1998 Q2	1.67	6.49	0	0	0	1
1998 Q3	−0.41	4.34	0	0	0	1
1998 Q4	−6.15	22.4	1	0	0	0
1999 Q1	−4.90	10.25	1	0	0	0
1999 Q2	−1.52	7.81	0	0	0	1
1999 Q3	−2.19	3.04	0	0	0	1
1999 Q4	−1.24	4.84	0	0	0	1

different hedge fund styles will exhibit different volatility exposures and hence will experience different outflow patterns.

Apart from flow data, we also need to identify the corresponding VIX changes for each quarter. In a slight deviation from the previous section, we do not look at the change in VIX for the quarter but, rather, we calculate the difference between the maximum VIX over the quarter and the VIX at the beginning of a quarter. In other words, we invest into long American digital calls rather than long European digital calls.[25] The difference in pricing will be subsumed in our parameter ε. This slight change is necessary as the VIX might spike within a quarter, creating large outflows, while its subsequent decline towards the event of the quarter will not reverse flows once money has left a hedge fund, ie, once investors have been scared. The VIX data is daily closes and come directly from CBOE, where they can be downloaded freely.[26]

Table 16.8 (*continued*)

			Case			
Date	**Flow (%)**	**ΔVIX**	**1**	**2**	**3**	**4**
2000 Q1	1.01	2.59	0	0	0	1
2000 Q2	−3.36	13.95	0	0	1	0
2000 Q3	1.08	1.31	0	0	0	1
2000 Q4	1.09	4.89	0	0	0	1
2001 Q1	4.56	4.2	0	0	0	1
2001 Q2	4.03	15.66	0	0	1	0
2001 Q3	3.55	11.42	0	0	1	0
2001 Q4	−5.74	11.51	1	0	0	0
2002 Q1	1.57	8.69	0	0	0	1
2002 Q2	2.20	1.31	0	0	0	1
2002 Q3	0.06	5.39	0	0	0	1
2002 Q4	−1.04	13.02	0	0	1	0
2003 Q1	2.55	5.54	0	0	0	1
2003 Q2	5.66	9.61	0	0	0	1
2003 Q3	6.70	0	0	0	0	1
2003 Q4	5.39	3.39	0	0	0	1
2004 Q1	12.07	5.3	0	0	0	1
2004 Q2	6.59	5.62	0	0	0	1
2004 Q3	1.70	6	0	0	0	1

Case 1: $\Delta\,\mathrm{AUM} \leqslant \Delta_{\mathrm{AUM}}$, $\Delta\mathrm{VIX} \geqslant \Delta_{\mathrm{VIX}}$; Case 2: $\Delta\,\mathrm{AUM} \leqslant \Delta_{\mathrm{AUM}}$, $\Delta\mathrm{VIX} < \Delta_{\mathrm{VIX}}$; Case 3: $\Delta\,\mathrm{AUM} > \Delta_{\mathrm{AUM}}$, $\Delta\mathrm{VIX} > \Delta_{\mathrm{VIX}}$; Case 4: $\Delta\,\mathrm{AUM} > \Delta_{\mathrm{AUM}}$, $\Delta\mathrm{VIX} < \Delta_{\mathrm{VIX}}$.

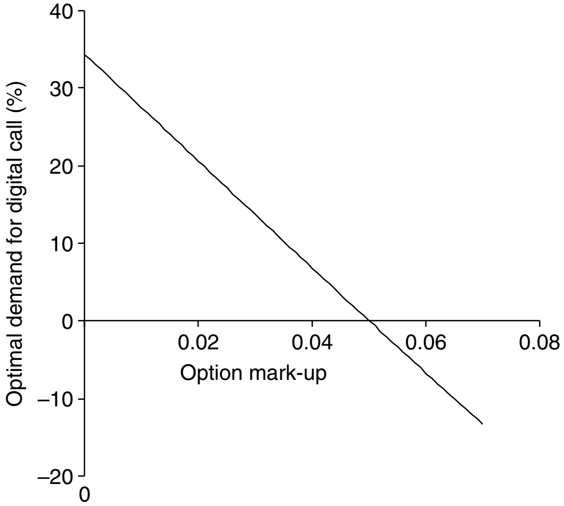
All the data is summarised in Table 16.8, where we have chosen $\Delta_{\mathrm{AUM}} = -3.514\%$. This represents the 90% confidence level on flows. In other words, portfolio outflows exceed -3.514% in only in one out of 10 quarters. This is consistent with our rationale for contingent hedging, ie, for hedging tail risk. The choice for $\Delta_{\mathrm{VIX}} = 10$ minimises Equation 16.16 and is equivalent in spirit to using an in-sample regression beta. For the data in Table 16.8 we can now fill the previously described contingency table (see Table 16.9), from which we can directly calculate all required inputs for the solution of Equation 16.17

$$\begin{aligned}\theta^{+} &= p(\Delta\,\mathrm{AUM} \leqslant \Delta_{\mathrm{AUM}}, \Delta\mathrm{VIX} \geqslant \Delta_{\mathrm{VIX}})\\ &\qquad \times [p(\Delta\,\mathrm{AUM} \leqslant \Delta_{\mathrm{AUM}}, \Delta\mathrm{VIX} \geqslant \Delta_{\mathrm{VIX}})\\ &\qquad\qquad + p(\Delta\,\mathrm{AUM} > \Delta_{\mathrm{AUM}}, \Delta\mathrm{VIX} \geqslant \Delta_{\mathrm{VIX}})]^{-1}\\ &= \frac{0.075}{0.2} = 0.375\end{aligned}$$

Table 16.9 Contingency table for hedge fund example

	Δ AUM $\leqslant \Delta_{\text{AUM}}$	Δ AUM $> \Delta_{\text{AUM}}$	
$\Delta\text{VIX} \geqslant \Delta_{\text{VIX}}$	0.075	0.125	0.2
$\Delta\text{VIX} < \Delta_{\text{VIX}}$	0.025	0.775	0.8
	0.1	0.9	1

Figure 16.1 Optimal hedging policy



$$\begin{aligned}\theta^{-} &= p(\Delta\,\text{AUM} \leqslant \Delta_{\text{AUM}}, \Delta\text{VIX} < \Delta_{\text{VIX}}) \\ &\qquad \times [p(\Delta\,\text{AUM} \leqslant \Delta_{\text{AUM}}, \Delta\text{VIX} < \Delta_{\text{VIX}}) \\ &\qquad\qquad + p(\Delta\,\text{AUM} > \Delta_{\text{AUM}}, \Delta\text{VIX} < \Delta_{\text{VIX}})]^{-1} \\ &= \frac{0.025}{0.8} = 0.03125 \\ v &= p(\Delta\text{VIX} \geqslant \Delta_{\text{VIX}}) = 0.2\end{aligned}$$

If we set $\lambda = 1$ and $f = 1$ we can calculate h^* as a function of ε, ie, how many American digital calls would the hedge fund manager optimally buy if the digital call price exceeds the real-world probability v (which it surely will) by $\varepsilon \geqslant 0$.

A graphical representation can be found in Figure 16.1. For our example, h^* approaches zero if the price for an American digital call (with strike 10 percentage points above the current VIX) exceeds

25%. The strategy works well with a hedging effectiveness of 75% (it pays out in 75% of all cases when a large outflow actually occurs)

$$p(\Delta \mathrm{VIX} \geqslant \Delta_{\mathrm{VIX}} \mid \Delta\,\mathrm{AUM} \leqslant \Delta_{\mathrm{AUM}}) = \frac{\theta^{+}v}{\theta^{+}v + \theta^{-}(1-v)} = \frac{0.075}{0.1} = \frac{3}{4} \tag{16.19}$$

However, it suffers from too many payouts that occur when they are actually not needed. This makes the option expensive for hedging purposes as it pays out in too many states (where protection is not needed) and creates costly variance in Equation 16.17.[27] Both effects reduce hedging demand. The results, however, confirm conventional hedge fund wisdom: buy long volatility when it is cheap.

APPENDIX A: APPROXIMATE DISTRIBUTION OF ASSET-BASED MANAGEMENT FEES

An asset-based fee at time $t+n, \tilde{f}_{t+n}$, is a random variable (characterised by $\sim$), that depends on the random realisation of the path of future assets under management $\tilde{A}_{t+i}$ for $i = 1, \dots, n$. More precisely, fees are calculated as a percentage θ (usually measured in basis points) on the average AUM over a given time period. We assume the AUM are calculated daily (which is the case for all retail funds with daily liquidity) and we will usually assume that $n = 250$, ie, we look at the distribution of annual fees with daily (almost continuous) averaging

$$\tilde{f}_{t+n} = \theta \cdot \left(\frac{1}{n} \sum_{i=1}^{n} \tilde{A}_{t+i} \right) \tag{16.20}$$

Ignoring future in- and outflows as well as active management returns, assets under management are tied to the evolution of benchmark returns $\tilde{A}_{t+i} = A_t \tilde{S}_{t+i}/S_t$, where $\tilde{S}_{t+i}/S_t$ denotes the benchmark return for the period from t to $t+n$. In other words, asset-management companies share the client's benchmark risks

$$\tilde{f}_{t+n} = \theta \cdot A_t \cdot \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\tilde{S}_{t+i}}{S_t} \right) \tag{16.21}$$

The trouble with Equation 16.21 is that even though $\tilde{S}_{t+i}/S_t$ is lognormal (with mean μ and variance σ^2) the sum of lognormal variables is not. However, Haug (2006) provides an approximate formula for a process with zero drift[28]

$$\ln\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{S}_{t+i}}{S_t}\right) \sim N\left(0, \sqrt{\ln\left(\frac{2\mathrm{e}^{\sigma^2}-2(1+\sigma^2)}{\sigma^4}\right)}\right) \tag{16.22}$$

where σ^2 denotes the variance of benchmark asset returns. Haug argues that

$$\ln\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{S}_{t+i}}{S_t}\right)$$

might very well be approximated by a normal distribution. The cumulative distribution for a lognormal variable is given by

$$P\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{S}_{t+i}}{S_t} \leqslant S\right) = \Phi\left(\ln(S)\Big/\sqrt{\ln\left(\frac{2\mathrm{e}^{\sigma^2}-2(1+\sigma^2)}{\sigma^4}\right)}\right) \tag{16.23}$$

where S denotes the average of rescaled (to 1) benchmark values and Φ stands for the cumulative density function of a standard normal. We can now easily calculate "fees-at-risk" (FaR) for alternative benchmark assets (ie, volatilities σ) and confidence level $1-\alpha$ from Equation 16.23 by solving for the required percentile. For $\Phi(z_\alpha) = 0.05$ we know that $z_\alpha = -1.64$. Given that expected returns are notoriously difficult to forecast, we assume that benchmark assets exhibit zero drift

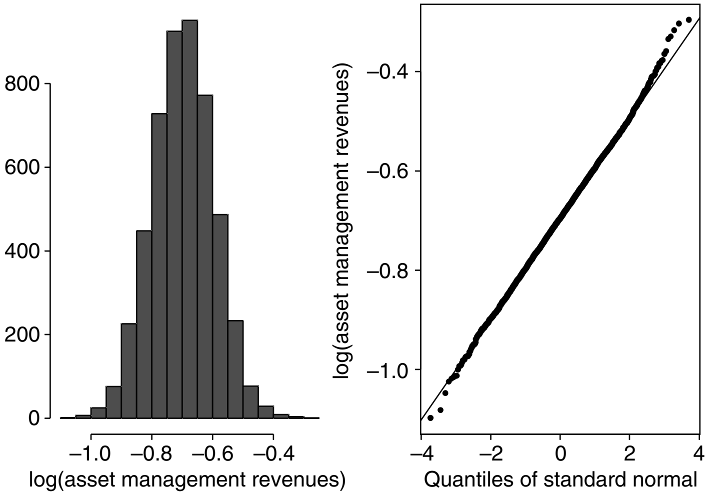
$$\mathrm{FaR}_{\alpha,t} = \theta \cdot A_t \cdot \exp\left\{-z_\alpha\sqrt{\ln\left(\frac{2\mathrm{e}^{\sigma^2}-2(1+\sigma^2)}{\sigma^4}\right)}\right\} \tag{16.24}$$

Also note that we can use Equations 16.21 and 16.22 to calculate the variance of asset-management fees. Applying again properties of the lognormal distribution, we get

$$\begin{aligned}\mathrm{var}(\tilde{f}_{t+n}) &= (\theta \cdot A_t)^2\,\mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{S}_{t+i}}{S_t}\right)\\ &= (\theta \cdot A_t)^2\left(\frac{2\mathrm{e}^{\sigma^2}-2(1+\sigma^2)}{\sigma^4}\right)\left(\frac{2\mathrm{e}^{\sigma^2}-2(1+\sigma^2)}{\sigma^4}-1\right)\end{aligned} \tag{16.25}$$

which is what we used in Equation 16.20. Note that for deriving Equation 16.25 we simply used the fact that for a lognormal random variable X with variance σ^2 we know that $\mathrm{var}(X) = \mathrm{e}^{\sigma^2}(\mathrm{e}^{\sigma^2}-1)$.

Figure 16.2 Histogram and quantile plot for bootstrapped log asset-management fees

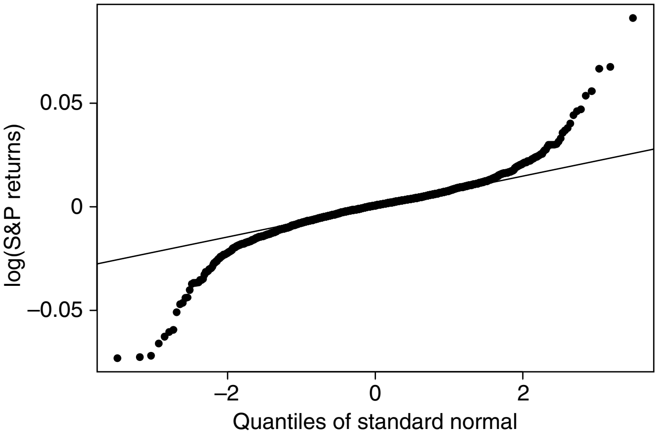


APPENDIX B: SIMULATION STUDY

We want to test whether $\ln(\tilde{f}_{t+n})$ is approximately normally distributed, ie, whether asset-management fees can be approximated as a lognormal random variable. If it can, a whole battery of approximations are available that could provide us with closed-form solutions for FaR or fee volatility. In order to perform this task, we bootstrap 5,000 annual time series for stock returns. To make the example realistic, we use real-world stock returns, ie, we use daily returns on the S&P 500 ranging from January 2000 to November 2008, to bootstrap from. It is well known that classic bootstrapping destroys dependence structures in a time series, as each draw is assumed to be drawn independently. Instead, we try to maintain some of the dependence structure with a straightforward modification. After each return draw we draw a second random variable from a uniform distribution between 0 and 1. If the draw falls below $q = \frac{1}{2}$, we take the neighbouring entry from the original data series; otherwise we continue drawing randomly from the original time series.

The likelihood of drawing two consecutive entries is then 0.5, the likelihood of three consecutive entries is 0.25 and so on, ie, the expected block length is 2. If q becomes larger, the expected block length increases. Figure 16.2 shows both a histogram and a quantile

Figure 16.3 Quantile plot for bootstrapped log S&P 500 returns



plot for $\ln(\tilde{f}_{t+n})$. Both graphs seem to confirm our assumption that asset-management fees can be approximated by a lognormal random variable. The QQ-plot shows a remarkably good fit, given that the underlying daily return series for the S&P 500 is far from normal, as we can see from Figure 16.3. Daily returns obviously exhibit large kurtosis.

This is a remarkable result (that also holds for larger q). However, while Figure 16.2 suggests a perfect fit, it is not. The term "approximately lognormal" still applies. Formal tests, like the Wilkinson–Shapiro test for normality of $\ln(\tilde{f}_{t+n})$ reject normality with a p value of around zero. Even though skewness (0.0142) and kurtosis (0.143) are small, they are still significant given the large number of observations. While this result is encouraging, we are unable to generalise for other time series, time periods or data frequencies, ie, "if in doubt, bootstrap".

APPENDIX C: MARKET BETA FOR ASSET-BASED FEES

Asset-based fees are based on average prices over a year, while market sensitivities are calculated using year-end prices, ignoring the "middle part" of the stock price path over a year. How will this affect estimated stock market sensitivities? How do the returns for the market portfolio and returns on average fund prices move together? We start from our definition of market sensitivity

$$\beta_{\text{fee}} = \text{cov}\left(\frac{n^{-1}\sum_{i=1}^{n} S_i^{\text{p}}}{S_0^{\text{p}}}, \frac{S_n^{\text{m}}}{S_0^{\text{m}}}\right) \Big/ \left(\text{var}\left(\frac{S_n^{\text{m}}}{S_0^{\text{m}}}\right)\right) \tag{16.26}$$

where S_i^{m} (S_i^{p}) denotes the market (portfolio) level at time i. Without loss of generality we set $S_0^{\text{m}} = S_0^{\text{p}} = 1$. Assume that portfolio and market prices follow

$$S_n^{\text{p}} = S_0^{\text{p}} + \sum_{i=1}^{n} \Delta S_i^{\text{p}} \quad \text{and} \quad S_n^{\text{m}} = S_0^{\text{m}} + \sum_{i=1}^{n} \Delta S_i^{\text{m}}$$

where $\text{var}(\Delta S^{\text{m}}) = \sigma_{\text{m}}^2 n$ such that

$$\text{var}(S_n^{\text{m}}) = \text{var}\left(\sum_{i=1}^{n} \Delta S_i^{\text{m}}\right) = n\,\text{var}(\Delta S) = \sigma_{\text{m}}^2 n^2 \tag{16.27}$$

We can now work out the covariance between average portfolio prices over n days and the final market price at day n

$$\begin{aligned}
&\text{cov}\left(\frac{1}{n}\sum_{i=1}^{n} S_i^{\text{p}}, S_n^{\text{m}}\right) \\
&\quad = \text{cov}\left(\frac{1}{n}\left(\underbrace{S_0^{\text{p}} + \Delta S_1^{\text{p}}}_{S_1^{\text{p}}} + \underbrace{S_0^{\text{p}} + \Delta S_1^{\text{p}} + \Delta S_2^{\text{p}}}_{S_2^{\text{p}}} + \cdots\right), S_0^{\text{m}} + \sum_{i=1}^{n} \Delta S_i^{\text{m}}\right)
\end{aligned} \tag{16.28}$$

Let us assume further that $\Delta S_i^{\text{p}} = \beta \Delta S_i^{\text{m}} + \varepsilon_i$

$$\begin{aligned}
\text{cov}\left(\frac{1}{n}\sum_{i=1}^{n} S_i^{\text{p}}, S_n^{\text{m}}\right) &= \text{var}(\beta \Delta S_1^{\text{m}}) + \frac{n-1}{n}\text{var}(\beta \Delta S_2^{\text{m}}) + \cdots \\
&\qquad + \frac{n-(n-1)}{n}\text{var}(\beta \Delta S_n^{\text{m}}) \\
&= (1 + \tfrac{1}{2}(n-1))\beta \sigma_{\text{m}}^2 n
\end{aligned} \tag{16.29}$$

Substitute Equations 16.29 and 16.28 into Equation 16.26 and we arrive at

$$\beta_{\text{rev}} = \frac{(1 + \frac{1}{2}(n-1))\beta \sigma_{\text{m}}^2 n}{\sigma_{\text{m}}^2 n^2} = \frac{\beta(1 + \frac{1}{2}(n-1))}{n} \overset{n\to\infty}{=} \tfrac{1}{2}\beta \tag{16.30}$$

Even though asset-based fees in the above example are determined by market returns, regression betas will equal half the market beta as a result of the averaging process.

[1] At a lower operational leverage of 50% (with the same total costs at the beginning of the year) profits would have still been up to US$36.25 million.

[2] However, it must be said that the fact that market timing is an activity that few managers engage in would let us conclude that beta timing is not regarded as a core competency that asset-management companies think they can earn economic profits from.

[3] Writing a chapter on this subject might look procyclical or like someone confusing hindsight knowledge with risk management. However, all the building blocks used in this chapter have been readily available for quite some time but it always takes a crisis before people start to listen.

[4] See Appendix A (on page 465) for a derivation of Equation 16.1 and its underlying assumptions together with an expression for FaR. Appendix B (on page 467) provides a brief simulation study on its approximating properties.

[5] Most asset managers are still owned by banks. However, a similar argument applies to stand-alone asset-management companies. Asset managers that show large losses are likely to experience larger redemptions on client concerns on their ability to keep and recruit key staff, etc.

[6] If in doubt, just recall what happened to CIOs of real estate asset-management firms in 2006 and 2007. During the property bubble, their management got highly paid for running a long beta business and often got promotions within a firm's hierarchy. When the bubble burst, they left the firm exposed. And of course the same applies to fixed-income managers that invested into credit versus a government bond benchmark up to 2007.

[7] All arguments used here have been available for decades in the corporate finance literature. A nice summary can be found in Doherty (2000). One application of the thoughts in this chapter is to the business of managing client money.

[8] In general QR provides an estimate of the quintile, q, of revenues, rev, as a linear function of our measure for market returns

$$r\,\widehat{\mathrm{rev}}_q \mid r = F_q^{-1}(q \mid r) = \alpha_q + \beta_q r$$

while an OLS regression takes the form $\widehat{\mathrm{rev}} = \alpha + \beta\hat{r}$. Note that $F_q^{-1}(q \mid r)$ is the inverse of the cumulative density function, conditional on our measure of market returns, r. In other words $F_q^{-1}(q \mid r)$ is the value at risk (conditional on r) at the $1 - q$ confidence level as a linear function of r, where the regression sensitivities change with q. We can think of β_q as the solution to

$$\beta_q = \arg\min_{\beta} \sum_{t=1}^{T} d_t q |\varepsilon_t| + (1 - d_t)(1 - q)|\varepsilon_t|$$

where $\varepsilon_t = \mathrm{rev}_t - \alpha - \beta x_t$ and $d_t = 1$ for $\varepsilon_t \geqslant 0$ and 0 otherwise. This compares to OLS as the solution to

$$\beta_{\mathrm{OLS}} = \arg\min_{\beta} \sum_{t=1}^{T} \varepsilon_t^2$$

[9] Note that this will also contain changes in asset mix over time as well as distributional strength and weakness.

[10] See Lewent and Kearney (1993) for a similar approach applied to currency risks.

[11] We use ANOVA to test using the variance of residuals for different conditional means.

[12] A replicating strategy would sell e^{-r}/n forward contracts for each of the n averaging points.

[13] In other words if

$$\bar{S} = S_t \sum_{i=1}^{n} \frac{\mathrm{e}^{ri/250}}{n}$$

[14] This example should be used for illustrative purposes only. As (all) asset managers manage various products with different benchmarks, we need to hedge each product separately. As this is a straightforward extension of what is presented in this chapter, we continue to present the "single product" case.

15 This choice is likely to underestimate the amount of beta exposure in revenue streams as the business mix of asset-management firms differ. It again makes our result very conservative, as we could easily find higher systematic equity exposure by tailoring equity market definitions to how AUM are invested in.

16 A purely fixed-income house is likely not have a large stock market sensitivity. In fact, the sensitivity might be negative if client inflows into fixed-income assets stall or even reverse in booming equity markets. None of the asset-management firms involved can be qualified as fixed-income houses.

17 This is similar to practitioners that weight regression coefficients based on their t values.

18 Alternatively, we could estimate Equation 16.6 using SUR and test the coefficient restrictions using a Wald test.

19 At the time of writing, Connor *et al* (2010) is the best source on portfolio risk asset management, but it entirely focuses on managing risks on behalf of clients instead of managing risk to the asset-management organisation.

20 See Baquero and Verbeek (2009) for a detailed study.

21 According to the author's private discussions with London hedge fund managers and sell side derivatives trading desks.

22 The set-up chosen is similar to Caballero and Panageas (2008), who derive optimal hedging demand within an FX reserve management context. The latter, however, focused on $\varepsilon = 0$ and hence on the variance term instead, which is somewhat odd given that $\varepsilon = 0$ assumes risk neutrality.

23 The necessary condition for a combined positive position in the digital option (hedging and risk minimisation demand) is given by

$$\theta^{+} > \left(\frac{\varepsilon(2\lambda - 1) + 2\lambda(1 - v)(v + \varepsilon)f}{2\lambda v(1 - (v + \varepsilon))f} \right)$$

where the factor in brackets exceeds 1 for $\varepsilon = 0$. Our set-up requires a larger differential between θ^{+} and θ^{-} than in Caballero and Panageas (2008) that only required $\theta^{+} > \theta^{-}$ for $\varepsilon = 0$.

24 Whether monthly, quarterly or annual redemptions are best to be hedged is an empirical question that the author does not intend to answer here. However, the author conjectures that annual hedges might be more effective by removing short term noise.

25 An American digital call will be exercised as soon as the barrier is hit.

26 See http://www.cboe.com/data/.

27 The variance penalty directly arises from the assumed mean variance objective.

28 We use Equation 16.22 for illustration given the community's obsession with closed-form solutions. The interested reader might explore various approximations for the distribution of

$$\ln\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{S}_{t+i}}{S_t}\right)$$

usually provided in the options pricing literature, as, for example, in Nelken (1996). However, given that log returns themselves are neither normal, uncorrelated nor independent all these expressions can only be seen as back-of-the-envelope shortcuts. Given the low (computational) costs of bootstrapping the small sample distribution, a simulation approach should always be the preferred route to a solution.

REFERENCES

Anson, M., 2002, "Symmetric Performance Measures and Asymmetric Trading Strategies", *Journal of Alternative Investments* 5(2), pp. 81–85.

Baltagi, B., 2005, *Econometric Analysis of Panel Data*, Third Edition (Chichester: John Wiley & Sons).

Baquero, G., and M. Verbeek, 2009, "A Portrait of Hedge Fund Investors: Flows, Performance and Smart Money", SSRN Working Paper.

Caballero, R., and S. Panageas, 2008, "Hedging Sudden Stops and Precautionary Contractions", *Journal of Development Economics* 85, pp. 28–57.

Chevalier, J., and G. Ellison, 1997, "Risk Taking by Mutual Funds as a Response to Incentives", *Journal of Political Economy* 105, pp. 1167–200.

Connor, G., L. R. Goldberg and R. A. Korajceck, 2010, *Portfolio Risk Management* (Princeton, NJ: Princeton University Press).

Doherty, N., 2000, *Integrated Risk Management* (New York: McGraw Hill).

Haug, E. G., 2006, *The Complete Guide to Option Pricing Formulas*, Second Edition (New York: McGraw-Hill).

Hull, J., 2007, *Risk Management and Financial Institutions* (Englewood Cliffs, NJ: Prentice Hall).

Lewent, J., and J. Kearney, 1993, "Identifying, Measuring and Hedging Currency Risk at Merck", in Donald Chew (ed) *The New Corporate Finance* (New York: McGraw-Hill).

Nelken, I., 1996, *The Handbook of Exotic Options* (New York: McGraw-Hill).

Ross, S., 2005, *Neoclassical Finance* (Princeton, NJ: Princeton University Press).

Sirri, E., and P. Tufano, 1998, "Costly Search and Mutual Fund Flows", *The Journal of Finance* 53, pp. 1589–622.

Verbeek, M., 2004, *Modern Econometrics* (Chichester: John Wiley & Sons).

Index

(page numbers in italic type relate to tables or figures)