Forecasting Expected Returns in the Financial Markets

Edited by

Stephen Satchell

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
84 Theobald's Road, London WC1X 8RR, UK
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
525B Street, Suite 1900, San Diego, California 92101-4495, USA

First edition 2007

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

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

ISBN: 978-0-7506-8321-0

For information on all Academic Press publications visit our web site at http://books.elsevier.com

Printed and bound in Great Britain

07 08 09 10 11 10 9 8 7 6 5 4 3 2 1

Contents

List of contributors

Robert Almgren
Departments of Mathematics and Computer Science, University of Toronto, Toronto, ON M5S, Canada

Neil Chriss
Department of Mathematics and Center for Financial Mathematics, University of Chicago, Chicago, IL 60637, USA

Theo Darsinos
Faculty of Economics, University of Cambridge, Cambridge CB2 1TQ, UK

Wayne Ferson
Marshall School of Business, University of Southern California, 701 Exposition Blvd, Los Angeles, CA

Edward Fishwick
Managing Director, Head of Equity Risk & Quantitative Analysis, BlackRock, 33 King William Street, London, EC4R 9AS, UK

Soosung Hwang
Reader in Finance, Cass Business School, 106 Bunhill Row, London EC1Y 8TZ

Thomas Idzorek
Director of Research, Ibbotson Associates, 225 North Michigan Avenue, Chicago, IL, USA

David Johnstone
School of Business, University of Sydney, NS 2006, Australia

Christos Koutsoyannis
Deputy Head of Quant Research, Old Mutual Asset Managers (UK) Ltd, 2 Lambeth Hill, London EC4P 4WR

Chensheng Lu
Faculty of Finance, Cass Business School, 106 Bunhill Row, London EC1Y9TZ, UK

Stephen Satchell
Reader in Financial Econometrics, Faculty of Economics, University of Cambridge, Cambridge CB2 1TQ, UK

Alan Scowcroft
Head of Equities Quantitive Research at UBS Warburg

Oliver Williams
PhD student, King's College, University of Cambridge, Cambridge, UK

Steve Wright
Fixed Income Department, UBS Investment Bank

Introduction

This book about forecasting expected returns covers a great deal of established topics in the forecasting literature, but also looks at a number of new ones. Wayne Ferson contributes a chapter on market efficiency and forecasting returns. Thomas Idzorek and Alan Scowcroft provide analyses of the Black–Litterman model; a methodology of great importance to practitioners in that it allows for the combination of analysts' views with equilibrium modelling.

Several chapters concern ranked returns. Neil Chriss and Robert Almgren provide an original and stimulating analysis of the use of rankings in portfolio construction. On the same topic, Steve Wright shows how ranking forecasts can be implemented in practice. Theo Darsinos has contributed two chapters concerned with forecasting option prices; this is an under-studied area.

Christos Koutsoyannis investigates how robust optimization can be used to improve forecasting returns. He also provides a thorough overview of the robust optimization literature. Soosung Hwang looks at the role of liquidity as an explanatory variable in forecasting equity returns.

Edward Fishwick and Oliver Williams contribute two chapters on the analysis of forecasting horizons – again, an area where very little work has been published.

Finally, David Johnstone and Oliver Williams present analyses of forecasting in a binomial world. Although this is a very simple framework, it nevertheless leads to important insights into such issues as to how forecasts can move markets.

Having put all this material together, I'm conscious that there is more that has not been included. I hope to rectify this at some later point.

Stephen Satchell

1 Market efficiency and forecasting

Wayne Ferson

1.1 Introduction

The interest in predicting stock prices or returns is probably as old as the markets themselves, and the literature on the subject is enormous. Fama (1970) reviews early work and provides some organizing principles. This chapter concentrates selectively on developments following Fama's review. In that review, Fama describes increasingly fine information sets in a way that is useful in organizing the discussion. Weak-form predictability uses the information in past stock prices. Semi-strong form predictability uses variables that are obviously publicly available, and strong form uses anything else. While there is a literature characterizing strong-form predictability (e.g. analyzing the profitability of corporate insider's trades), this chapter concentrates on the first two categories of information.

For a while, predicting the future price or value (price plus dividends) of a stock was thought to be easy. Early studies, reviewed by Fama (1970), concluded that a martingale or random walk was a good model for stock prices, values or their logarithms. Thus, the best forecast of the future price was the current price. However, predicting price or value changes, and thus rates of return, is more challenging and controversial. The current financial economics literature reflects two often-competing views about predictability in stock returns. The first argues that any predictability represents exploitable inefficiencies in the way capital markets function. The second view argues that predictability is a natural outcome of an efficient capital market.

The exploitable inefficiencies view of return predictability argues that, in an efficient market, traders would bid up the prices of stocks with predictably high returns, thus lowering their return and removing any predictability at the new price (see, for example, Friedman, 1953; Samuelson, 1965). However, market frictions or human imperfections are assumed to impede such price-correcting, or 'arbitrage' trading. Predictable patterns can thus emerge when there are important market imperfections, like trading costs, taxes, or information costs, or important human imperfections in processing or responding to information, as studied in behavioural finance. These predictable patterns are thought to be exploitable, in the sense that an investor who could avoid the friction or cognitive imperfection could profit from the predictability at the expense of other traders.

The 'efficient markets' view of predictability was described by Fama (1970). According to this view, returns may be predictable if required expected returns vary over time in association with changing interest rates, risk or investors' risk-aversion. If required expected returns vary over time there may be no abnormal trading profits, and thus no incentive to exploit the predictability. Predictability may therefore be expected in an efficient

capital market. The return is written as $R = E(R|\Omega) + u$, where Ω is the information at the beginning of the period and u is the unexpected return. Since $E(u|\Omega) = 0$, the unexpected return cannot be predicted ahead of time. Thus, predictability, in the 'efficient markets' view, rests on systematic variation through time in the expected return. Modelling and testing for this variation is the focus of the conditional asset pricing literature (see reviews by Ferson, 1995; Cochrane, 2005).

While this chapter focuses on return predictability, not all of the predictability associated with stock prices involves predicting the levels of returns. A large literature models predictable second moments of returns (e.g. using ARCH and GARCH-type models (see Engle, 2004) or other stochastic volatility models). Predictability studies have also examined the third moments (see, for example, Harvey and Siddique, 2001).

1.2 A modern view of market efficiency and predictability

As described by Fama (1970), any empirical analysis of stock return predictability or the market's informational efficiency involves a 'joint hypothesis'. There must be an hypothesis about the model for equilibrium expected returns, and also an hypothesis about the informational efficiency of the markets. These can be easily described using a modern representation for asset pricing models.

Most of the asset pricing models of financial economics can be described as versions of equation (1.1):

$$E\{m_t(1 + r_t)|\Omega_{t-1}\} = 1 \tag{1.1}$$

where Ω_{t-1} is the information set of economic agents at the beginning of the period and r_t is the rate of return of a financial asset. The scalar random variable m_t is the stochastic discount factor. Different models imply different stochastic discount factors, and the stochastic discount factor should price all of the assets in the model through equation (1.1).

The joint hypothesis of stock return predictability and market efficiency tests may now be described. The assumption of a model of market equilibrium amounts to a specification for m_t. For example, the Capital Asset Pricing Model of Sharpe (1964) implies that m_t is a linear function of the market portfolio return (see, for example, Dybvig and Ingersoll, 1982). Assume that the analyst uses the lagged variables, Z_{t-1}, to predict stock returns. The hypothesis of informational efficiency is simply the statement that Z_{t-1} is contained in Ω_{t-1}. For example, weak-form efficiency says that past stock prices are in Ω_{t-1}, while semi-strong form efficiency says that other publicly available variables are in Ω_{t-1}.

The martingale model for stock values follows as a special case of this modern view. If we assume that m_t is a constant over time (as implied by risk neutral agents with fixed time discounting), then equation (1.1) implies that $E\{r_t|\Omega_{t-1}\}$ is a constant over time. Since $r_t = E(r_t|\Omega_{t-1}\} + u_t$, and u_t is unpredictable, it follows that the returns r_t cannot be predicted by any information in Ω_{t-1}.

1.3 Weak-form predictability

Much of the literature on weak-form predictability can be characterized through an autoregression. Let R_t be the continuously compounded rate of return over the shortest measurement interval ending at time t. Let $r(t, t+H) = \Sigma_{j=1,\ldots,H} R_{t+j}$. Then,

$$r(t, t+H) = a_H + \rho_H r(t-H, t) + \varepsilon(t, t+H) \tag{1.2}$$

is the autoregression and H is the return horizon. Studies can be grouped according to the return horizon.[1]

Many studies measure small but statistically significant serial dependence in daily or intra-daily stock return data. Serial dependence in daily returns can arise from end-of-day price quotes that fluctuate between bid and ask (Roll, 1984), or from non-synchronous trading of the stocks in an index (see, for example, Fisher, 1966; Scholes and Williams, 1977). These effects do not represent predictability that can be exploited with any feasible trading strategy. Spurious predictability due to such data problems should clearly not be attributed to time-variation in the expected discount rate for stocks. On the other hand, much of the literature on predictability allows that high frequency serial dependence may reflect changing conditional means. For example, Lo and MacKinlay (1988) and Conrad and Kaul (1988) model expected returns within the month as following an autoregressive process.

Conrad and Kaul (1988, 1989) studied serial dependence in weekly stock returns. They point out that if the expected returns, $E(R|\Omega)$, follow an autoregressive process, the actual returns would be described by the sum of an autoregressive process and a white noise, and thus follow an ARMA process. The autoregressive and moving average coefficients would be expected to have the opposite signs: If current expected returns increase, it may signal that future expected returns are higher, but stock prices may fall in the short run because the future cash flows are discounted at a new, higher rate. The two effects, offset and returns, could have small autocorrelations. Estimating ARMA models, they found that the autoregressive coefficient for weekly returns on stock portfolios are positive, near 0.5, and can explain up to 25% of the variation in the returns on a portfolio of small-firm stocks.

Even with weekly returns, however, some of the measured predictability can reflect non-synchronous trading effects. Lo and MacKinlay (1990) and Muthaswamy (1988) use statistical models that attempt to separate out the various effects in measured portfolio returns. Boudoukh *et al.* (1994) use stock index futures contracts, which are not subject to non-synchronous trading, and find little evidence for predictability at a weekly frequency.

Much of the literature on weak-form predictability studies broad stock market indexes or portfolios of stocks, grouped according to the market capitalization (size) or other characteristics of the firms. However, another significant stream of the literature studies relative predictability. Stocks have relative predictability if the future returns of one group of stocks are predictably higher than the returns of another group. Thus, if a trader could

[1] An alternative to the autoregression is the Variance Ratio statistic, $\text{Var}\{r(t, t+H)\}/HVar(R_t)$, proposed by Working (1949) and studied for stock returns by Lo and MacKinlay (1988, 1989) and others. Cochrane (1988) shows that the variance ratio is a function of the autocorrelation in returns. Kaul (1996) provides an analysis of various statistics that have been used to evaluate weak-form predictability, showing how they can be viewed as combinations of autocorrelations at different lags, with different weights assigned to the lags.

buy the stocks in the high-return group and sell short the stocks in the low-return group, the trader could profit even if both groups were to go up (or down). In a weak-form version of relative predictability, past stock prices or returns are used to form the groups. If past winner (loser) stocks have predictably higher (lower) returns, we have continuation or 'momentum'. If past winner stocks can be predicted to have lower future returns, we have 'reversals'. Relative predictability can be evaluated by viewing equation (1.2) as a cross-sectional regression – an approach taken by Jegadeesh (1990). Lehman (1990) finds some evidence for reversals in the weekly returns of US stocks.

Monthly returns are commonly used in the literature that tests asset pricing models. At this frequency, the evidence on weak-form predictability is relatively sparse. Jegadeesh (1990) finds some evidence for reversals at a monthly return frequency. Ferson *et al.* (2005) make indirect inferences about the time-variation in monthly expected stock returns by comparing the unconditional sample variances of monthly returns with estimates of expected conditional variances. The key is a sum-of-squares decomposition: $Var(R)\} = E\{Var(R|\Omega)\} + Var\{E(R|\Omega)\}$, where $E(.|\Omega)$ and $Var(.|\Omega)$ are the conditional mean and variance, and $Var\{.\}$ and $E\{.\}$, without the conditioning notation, are the unconditional moments. The interesting term is $Var\{E(R|\Omega)\}$; that is, the amount of variation through time in conditionally expected stock returns. This quantity is inferred by subtracting estimates of the expected conditional variance, $E\{Var(R|\Omega)\} = E\{[R - E(R|\Omega)]^2\}$, from estimates of the unconditional variance. The expected conditional variance is estimated following Merton (1980), who showed that while the mean of a stock return is hard to estimate, it is almost irrelevant for estimating the conditional variance, when the time between observations is short. Using high-frequency returns to estimate the conditional variance for each month, then subtracting its average from the monthly unconditional variance, the difference – according to the decomposition – is the variance of the monthly conditional mean.

Ferson *et al.* (2005) find that while historical data prior to 1962 suggests economically significant weak-form predictability in monthly stock market returns, there is little evidence of weak-form predictability for monthly returns in modern data. In particular, the evidence for the period after 1992 suggests that any weak-form predictability in the stock market as a whole has vanished. At the same time, a simulation study shows that the indirect tests have the power to detect even modest amounts of predictability.

Jegadeesh and Titman (1993) find that relatively high-past-return stocks tend to repeat their performance over 3- to 12-month horizons. They study US data for 1927–1989, but focus on the 1965–89 period. The magnitude of the effect is striking. The top 20% winner stocks over the last 6 months can outperform the loser stocks by about 1% per month for the next 6 months. This momentum effect has spawned a huge subsequent literature which is largely supportive of the momentum effect, but which has not reached a consensus about its causes. The efficient markets view of predictability suggests that momentum trading strategies should be subject to greater risk exposures which justify their high returns. Most efforts at explaining the effect by risk adjustments have failed.[2]

The momentum effect has inspired a number of behavioural models, suggesting that momentum may occur because markets under-react to news in the pricing of stocks.

[2]There are some partial successes. For example, Ang *et al.* (2001) associate some of the momentum strategy profits with high exposure to 'downside risk' – that is, the covariance with market returns when the market return is negative.

For example, one argument (Daniel *et al.*, 1998) is that traders have 'biased self attribution', meaning that they think their private information is better than it really is. As a result they do not react fully to public news about the value of stocks, so the news takes time to get impounded in market prices, resulting in momentum. In another argument, traders suffer a 'disposition effect', implying that they tend to hold on to their losing stocks longer than they should, which can lead to momentum (Grinblatt and Han, 2003). These arguments suggest that traders who can avoid these cognitive biases may profit from momentum trading strategies. However, Lesmond *et al.* (2004) and Korajczyk and Sadka (2004) measure the trading costs of momentum strategies and conclude that the apparent excess returns to the strategies are consumed by trading costs.

Perhaps the most controversial evidence of weak-form predictability involves long-horizon returns. Fama and French (1988) use autoregressions like equation (1.2) to study predictability in portfolio returns, measured over 1-month to multi-year horizons. They find U-shaped patterns in the autocorrelations as a function of the horizon, with negative serial dependence, or mean reversion, at 4- to 5-year horizons. Mean reversion can be consistent with either view of predictability. If expected returns are stationary (reverting to a constant unconditional mean) but time-varying, mean reversion can occur in an efficient market. Mean reversion would also be expected if stock values depart temporarily from the fundamental, or correct, prices, but are drawn back to that level. The evidence for weak-form predictability in long-horizon returns is subject to a number of criticisms on statistical grounds, as described below.

DeBondt and Thaler (1985) find that past high-return stocks perform poorly over the next 5 years, and *vice versa*, a form of relative predictability. They interpret reversals in long-horizon relative returns as evidence that the market over-reacts to news about stock values, and then eventually corrects the mistake. The reversal effect was shown to occur mainly in the month of January, by Zarowin (1990) and Grinblatt and Moskowitz (2003), which is interpreted as related to 'tax loss selling'. In this story, investors sell loser stocks at the end of the year for tax reasons, thus depressing their prices, and buy them back in the new year, subsequently raising their prices. McLean (2006) finds that reversals are concentrated in stocks with high idiosyncratic risks, which is thought to present a deterrent to arbitrage traders who might otherwise correct temporary errors in the market prices.

Like momentum, behavioural models attempt to explain reversals as the result of cognitive biases. Models of Barberis *et al.* (1998), Daniel *et al.* (1998) and Hong and Stein (1999) argue that both short-run momentum and long-term reversals can reflect biases in under- and over-reacting to news about stock values. Research in this area continues, and it's fair to say that the jury is still out on the issue of weak-form predictability in long-horizon returns.

1.4 Semi-strong form predictability

Studies of semi-strong form predictability can be described with the regression:

$$r(t, t+H) = \alpha_H + \beta_{H'} Z_t + v(t, t+H) \tag{1.3}$$

where Z_t is a vector of variables that are publicly available by time t. Many predictor variables have been analyzed in published studies, and it is useful to group them into

categories. The first category of predictor variables comprises 'valuation ratios', which are measures of cash flows divided by the stock price. Keim and Stambaugh (1986) use a constant numerator in the ratio and 'detrend' the price. Rozeff (1984), Campbell and Shiller (1988) and Fama and French (1989) use dividend/price ratios, Pontiff and Schall (1998) and Kothari and Shanken (1997) use the book value of equity divided by price. Boudoukh *et al.* (2004) and Lei (2006) add share repurchases and other non-cash payouts, respectively, to the dividend measure. Lettau and Ludvigson (2001) propose a macroeconomic variation on the valuation ratio: aggregate consumption divided by a measure of aggregate wealth. All of these studies find the regression coefficients β_H to be significant.

Malkiel (2004) reviews a valuation ratio approach that he calls the 'Federal Reserve Model.' Here, the market price/earnings ratio is empirically modelled as a function of producer prices, Treasury yields and other variables, and the difference between the model's output and the ratios observed in the market are used to predict the market's direction. (Malkiel finds that the model does not outperform a buy-and-hold strategy.)

Rozeff (1984) and Berk (1995) argue that valuation ratios should generally predict stock returns. Consider the simplest model of a stock price, P, as the discounted value of a fixed flow of expected future cash flows or dividends: $P = c/R$, where c is the expected cash flow and R is the expected rate of return. Then, $R = c/P$, and the dividend price ratio is the expected return of the stock. If predictability is attributed to the expected return, as in the efficient markets view, then a valuation ratio should be a good predictor variable.

Predictability of stock returns with valuation ratios is also related to the expected growth rates of future dividends or cash flows. Consider the Gordon (1962) constant-growth model for a stock price: $P = c/(R - g)$, where g is the future growth rate. Then, $c/P = R - g$. This suggests that if dividend/price ratios vary, either across stocks or over time, then expected returns should vary, and/or expected cash flow growth rates should vary, and the dividend/price ratio should be able to predict one or the other. Campbell and Shiller (1988) show that the intuition from this example holds to a good approximation in more general discounted cash flow models, where the growth rates and expected returns are not held fixed over time. They find that market dividend/price ratios do not significantly predict future cash flow growth rates. Cochrane (2006) uses this result to re-evaluate the empirical evidence for stock return predictability using dividend/price ratios. He essentially argues that if you know that the dividend/price ratio does not forecast future cash flow growth, then it must forecast future stock returns.

Studies of semi-strong form predictability in stock index returns typically report regressions with small R-squares, as the fraction of the variance in returns that can be predicted with the lagged variables is small – say 10–15% or less for monthly to annual return horizons. The R-squares are larger for longer-horizon returns – up to 40% or more for 4- to 5-year horizons. This is interpreted as the result of expected returns that are more persistent than returns themselves, as would be expected if returns are expected returns plus noise. Thus, the variance of the sum of the expected returns accumulates with longer horizons faster than the variance of the sum of the returns, and the R-squares increase with the horizon (see, for example, Fama and French, 1989). However, small R-squares can mask economically important variation in the expected returns.

Stocks are long 'duration' assets, so a small change in the expected return can lead to a large change in the asset value. To illustrate, consider another example using the Gordon

model, where the dividend is $c = kE$, E is the earnings and k is the dividend payout ratio. The price/earnings ratio $P/E = 15$, the payout ratio $k = 0.6$, and the expected growth rate $g = 3\%$. The expected return $R = 7\%$. Suppose there is a shock to the expected return, *ceteris paribus*. A change of 1% in R leads to approximately a 20% change in the asset value. Of course, this overstates the effect, to the extent that a shock that changes the required return also changes the expected future cash flows.

As the example suggests, small changes in expected returns can produce economically significant changes in asset values. Consistent with this argument, studies such as Kandel and Stambaugh (1996), Campbell and Viceira (2002) and Fleming *et al.* (2001) show that optimal portfolio decisions can be affected to an economically significant degree by return predictability, even when the amount of predictability, as measured by R-squared, is small.

The second category of semi-strong form predictor variables for stock returns includes calendar and seasonal effects. The list of effects that have been related to stock returns and the list of studies are too long to cite here (see Haugen and Lakonishok (1988) and Schwert (2003) for reviews). Some examples include the season (winter versus summer), the month of the year (especially, high returns in January), the time of the month (first versus last half), holidays, the day of the week (low returns on Mondays), the time of the day, the amount of sunlight (as in seasonal affective disorder) and even the frequency of geomagnetic storms.

The third category of predictor variables in equation (1.3) is a catch-all, 'other' variables. Prominent among these are bond yields and yield spreads. Fama and Schwert (1977) were among the first to observe that the level of short-term Treasury yields predicts returns in equation (1.3) with a negative coefficient. They interpreted the short-term yield as a measure of expected inflation. Ferson (1989) argues that the regressions imply that the systematic risk of stocks that determines the expected returns must vary over time with changes in interest rates. Keim and Stambaugh (1986) study the yield spreads of low-quality over high-quality bonds, and find predictive ability for stock returns, and Campbell (1987) studies a number of yield spreads in shorter-term Treasury securities. Fama and French (1989) assemble a list of variables from studies in the 1980s and describe their relations with US business cycles.

Another interesting set of predictor variables in equation (1.3) includes measures of the conditional variance or volatility of stock returns. Merton (1980) shows that a simple version of his Intertemporal Asset Pricing Model implies that expected returns on the aggregate stock market should be positively related to the conditional variance of the market returns; that is, there should be a positive risk-return tradeoff for the market as a whole. Sharpe's (1964) Capital Asset Pricing Model also makes this prediction. Early studies that tried to predict market returns with predetermined market volatility measures found mixed results – weak or even negative β_H coefficients (e.g. French *et al.*, 1987; Breen *et al.*, 1993). Scruggs (1998) showed that if additional risk factors as suggested by Merton's (1973) model were included, the partial coefficients became positive as predicted by the theory.

Of course, many other semi-strong form predictor variables have been proposed and more will doubtless be proposed in the future. Some recent variables include the fraction of equity issues in new issues of corporate securities (Baker and Wurgler, 2000), firms' investment plans (Lamont, 2000), the average 'idiosyncratic' or firm-specific component of past return volatility (Goyal and Santa Clara, 2003), the level of corporate cash holdings

(Greenwood, 2004), the aggregate rate of dividend initiation (Baker and Wurgler, 2000), share issuance (Pontiff and Woodgate, 2006), and the political party currently in office (Santa Clara and Valkanov, 2003).

1.5 Methodological issues

Even though the regressions in equations (1.2) and (1.3) seem pretty straightforward, interpreting the predictability evidence for stock returns based on these regressions is not. It can be argued that one of the greatest contributions of the literature on stock return predictability is the methodological lessons it has taught researchers in the field. Perhaps the most difficult issues involve selection bias and data mining. Additional issues that have been addressed in the literature include small sample biases in the coefficients, standard error estimation, multiple comparisons, efficient estimation, regime shifts, spurious regression and the interactions among these effects.

Selection bias and data mining are serious concerns. Data mining refers to sifting through the data in search of predictive or associative patterns. There are two kinds of data mining. Sophisticated data mining accounts for the number of searches undertaken when evaluating the statistical significance of the finding (see, for example, White, 2000). This is important, because if 100 independent variables are examined, we expect to find five that are 'significant' at the 5% level, even if there is no predictive relation. Naive data mining does not account for the number of searches. The big problem, given the strong interest in predicting stock returns among academics and practitioners, and the many studies using the same data, is that it is difficult to account for the number of searches. There probably have been at least as many regressions run using the Center for Research in Security Prices (CRSP) database as there are numbers in the database. Compounding this problem are various selection biases. Perhaps most difficult is the fact that only 'significant' results are circulated and published in academic papers. No one knows how many insignificant regressions were run before those results were found.

A reasonable response to these concerns is to see if the predictive relations hold out-of-sample. This kind of evidence is mixed. Some studies find support for predictability in step-ahead or out-of-sample exercises (for example, Fama and French, 1989; Pesaran and Timmerman, 1995). Semi-strong form variables show some ability to predict returns outside of the US data where they were originally studied (for example, Harvey, 1991; Ferson and Harvey, 1993, 1999; Solnik, 1993). Other studies conclude that predictability using many of the semi-strong form variables does not hold outside of the original samples (for example, Goyal and Welch, 2003, 2004; Simin, 2006). Even this evidence is difficult to interpret, because a variable could have real predictive power yet still fail to outperform a naive benchmark when predicting out of sample (Campbell and Thompson, 2005; Hjalmarsson, 2006).

A large literature has addressed statistical issues in predictive regressions. Boudoukh and Richardson (1994) provide an insightful review. From the perspective of testing the null hypothesis of no predictability, we are interested in whether β_H is zero. The coefficient is proportional to the covariance, $Cov\{r(t, t+H); Z_t\}$. The literature has employed three basic approaches to estimating the covariance. The first is to run the regression in equation (1.3) with overlapping data on $r(t, t+H)$, observed each period t,

usually 1 month (see, for example, Fama and French, 1989). Given overlapping data, this approach uses all of the data, as opposed to a sampling scheme, which measures non-overlapping returns every H months, and should therefore be more efficient than the sampling scheme. However, the overlap induces autocorrelation in the error terms in the form of an $H-1$ order moving average process. To conduct inference about β_H, it is necessary to estimate the coefficients and the standard errors without bias in the presence of the moving average error terms. This is complicated by the evidence that stock return data are conditionally heteroscedastic.

When $H > 1$ and several horizons are examined together, the issue of multiple comparisons arises. If 20 independent horizons are examined, it is expected that there will be one 'significant' t-statistic found at the 5% level when the null hypothesis of no predictability is true. The slope coefficients for the different horizons are correlated, which complicates the inference. Richardson (1993) shows this implies that the U-shaped patterns in the autocorrelations across return horizons, observed by Fama and French (1988), are likely to be observed by chance. Boudoukh *et al.* (2005) argue that high correlation of test statistics across the return horizons renders much of the evidence for semi-strong form long-horizon predictability suspect.

Even when $H = 1$ and there is no overlap, correlation in the error terms can lead to finite sample biases in estimates of the slope. Stambaugh (1999) studies one such bias that arises because the regressor is stochastic and its future values are correlated with the error term in the regression. For example, if Z_t is a dividend/price ratio, then shocks to the dividend price ratio at time $t+1$ are related to stock returns at time $t+1$ through the stock price. Stambaugh provides corrections for this bias, which he shows is related to the bias in a sample autocorrelation coefficient, as derived by Kendal (1954). Amihud and Horvitz (2004) explore solutions for this bias in a multiple regression setting.

A second approach to estimating $Cov\{r(t, t+H); Z_t\} = Cov\{\Sigma_{j=1,\ldots,H}\ R_{t+j}; Z_t\}$ is to recognize that, if the variables are covariance stationary, then $Cov\{\Sigma_{j=1,\ldots,H}\ R_{t+j}; Z_t\} = Cov\{R_t; \Sigma_{j=1,\ldots,H}\ Z_{t-j}\}$. This suggests using the horizon $H = 1$ on the left-hand side of equation (1.3), and replacing Z_t on the right-hand side with $\Sigma_{j=1,\ldots,H} Z_{t-j}$. This is the approach taken by Hodrick (1992) and Cochrane (1988). In this approach, the error terms in the regression are not overlapping and there is no induced moving average structure to account for. There may be efficiency gains compared with the first approach, but these depend on the stochastic process that drives the variables.

A third approach to estimating the predictive regression coefficient is to model the single period data $\{R_t,\ Z_{t-1}\}$ using a vector autoregression, and then infer the value of the long-horizon coefficient β_H from the autoregression parameters. This is the approach taken by Kandel and Stambaugh (1990), Campbell (1993), and others. This can be efficient if the vector autoregression is correctly specified, but is subject to error if the autoregression is not correctly specified.

A potential issue with all of the approaches to predictive regressions is spurious regression bias. Spurious regression is studied by Yule (1926) and Granger and Newbold (1974), who warn that empirical relations may be found between the levels of trending time series that are actually independent. For example, given two independent random walks, it is likely that a regression of one on the other will produce a 'significant' slope coefficient, evaluated by the usual t-statistics. In equation (1.3) the dependent variables are stock returns, which are not highly persistent. However, recall that the returns may be considered the sum of the expected return, plus unpredictable noise. If the expected returns are

persistent, even if stationary time series, there is still a risk of spurious regression in finite samples. Because the unpredictable noise represents a substantial portion of the variance of stock returns, spurious regression effects, will differ for stock returns, from those in the classical setting of Granger and Newbold. This version of the spurious regression problem has received some attention in recent econometric studies, but more attention to the problem is probably due.

The interaction among the various statistical issues with predictive regressions has received relatively little research to date, and I would expect this to be an active field in the future. Ferson *et al.* (2003) study the interaction between data mining and spurious regression effects. They find that data mining for predictor variables interacts with spurious regression bias in equation (1.3). The two effects reinforce each other, because more highly persistent series are more likely to be found significant in the search for predictor variables. Simulations suggest that many of the regressions in the literature, based on individual predictor variables, may be spurious. Powell *et al.* (2006) extend the analysis of the interaction between data mining and spurious regression to conclude that recent studies of presidential regimes (Santa-Clara and Valkanov, 2003), the 'Halloween Indicator' (Bouman and Jacobsen, 2002) and business cycle effects in momentum (Chordia and Shivakumar, 2002) appear insignificant in view of their combined effects.

1.6 Perspective

Does the current body of evidence lead to the conclusion that there actually is predictability in stock returns? I think there are good reasons to be sceptical of predictability and good reasons to believe in predictability.

Why should we be sceptical? First, the logic of the old random walk = efficient markets literature is compelling to many. In that view, traders would bid up the prices of stocks with predictably high returns, thus lowering their returns and removing any predictability at the new price. Furthermore, data mining and selection bias conspire to make us see predictable patterns where none may exist. There are many choices for the return horizon, H. It need not be the same on both sides of the regression in equation (1.2). Jegadeesh (1990) provides a statistical analysis of the choice of horizons in this context. However, it is the number of choices that leads to scepticism.

For another example, studies of relative predictability, such as momentum, have sliced and diced common stocks into portfolios based on many characteristics of the data, at which point the effect often retreats into subsets of stocks, subperiods of time, phases of the business cycle or other parts of the data. Studies find momentum to be concentrated according to industries, the size of the firm (more momentum in large stocks), the price of the share (more when the price is above $5 per share), etc. The effect does not appear prior to 1940, appears stronger after 1968, and appears stronger during economic expansions than contractions.

The evidence for long-term return reversals has a similar problem. Reversals have been found to be concentrated in small stocks, low-priced stocks, the month of January, high idiosyncratic risk stocks, and to be more pronounced in earlier samples than in more recent data. For each approach to slicing and dicing the data there is a clever story. That is not bad. By digging into subsamples it should be possible in principle, to isolate what

is driving an effect. But many of the patterns, especially those documented in the weak-form predictability literature, appear to be sample-specific. Finding results that vary with the sample period stimulates research featuring structural breaks and regime shifts. This reader is left more with concerns about naive data mining, improper multiple comparisons and statistical issues than with an understanding of why and where these weak-form effects occur systematically.

Predictive regressions are subject to a host of statistical issues. There are finite sample biases and problems associated with structural breaks and regime shifts. It is hard to get reliable standard errors for the regressions. There are potential spurious regression problems. And, as we are now beginning to understand, these effects can interact with each other. If semi-strong form predictability is spurious, as a result of statistical bias and naive data mining, we would expect predictor variables to appear in the empirical literature, then fail to work with fresh data. To some extent, the literature has evolved in this way.

There are also many good reasons to believe that stock returns are predictable. First of all, theory suggests that some amount of predictability is likely. If expected returns vary over time with some degree of persistence, predictability is expected. Most people find it easy to believe that expected stock returns and risks might be different coming out of a recession, for example, from going into one, and the predictability evidence tends to confirm such commonsense patterns. Studies find that predictability using lagged variables is largely explained by asset pricing models with multiple risk factors, if they allow the premiums associated with those risks to vary over time (see, for example, Ferson and Harvey, 1991; Ferson and Korajczyk, 1995; Avramov and Chordia, 2006). The behavioural models of predictability are compelling to many, as we see ourselves making the same cognitive errors as made by the agents in those models.

The momentum effect has been found to hold 'out of sample', relative to the original study of Jegadeesh and Titman (1993). Jegadeesh and Titman (2001) find momentum in data for 1990–1998. Momentum is found in stock markets outside of the US by Rouwehorst (1998) and Chui *et al.* (2000), among others.

The evidence of semi-strong form predictability has also survived a number of out-of-sample tests, working in other countries and over different time periods. Many of the variables identified as predictors for stock returns also seem to have some predictive power for other types of securities, such as bonds and futures, and also for the growth rates in 'fundamental' macroeconomic data. The evidence for semi-strong form predictability has survived corrections for a host of statistical and data problems.

Some studies that find semi-strong form stock market predictability, measured directly using lagged variables, has weakened in recent samples. It may be that the predictability was never really there, or that it was 'real' when first publicized, but diminished as traders attempted to exploit it. Ferson *et al.* (2005) examine semi-strong form predictability by regressing individual stocks on firm-specific predictors, then measuring the average covariances of the fitted values. It can be shown that the variance of the expected return on a large portfolio is approximately the average covariance. They find no evidence that predictability, measured in this way, is weaker in recent subperiods. As the firm-specific predictors have not been examined extensively in the literature, they may be less subject to naive data mining biases. Campbell and Thompson (2005), using step-ahead tests, also find that semi-strong form predictability holds up in recent data.

1.7 Conclusion

The issue of predictability in stock returns has important and broad economic implications. For example, it relates to the efficiency of capital markets in allocating resources to their highest valued uses. However, the interpretation of predictability, and the evidence for its very existence, remains controversial. This review of the literature finds the evidence for weak-form predictability (using the information in past stock prices) to be more fragile and less compelling than the evidence for semi-strong form predictability (using publicly available information more generally).

For the field of financial economics and asset pricing in particular, allowing for predictability through time-variation in expected returns, risk measures and volatility has been one of the most significant developments of the past two decades. Such conditional asset pricing models have provided a rich setting for the study of the dynamic behaviour of asset markets. For example, Conditional Performance Evaluation is the application of these models to the problem of evaluating the performance of portfolio managers. Models that allow for time-varying conditional moments produce different inferences about performance than do the traditional measures that do not allow for predictability (e.g. Ferson and Schadt, 1996), and they have influenced both academic views and professional investment practice. Research on predictability has stimulated numerous advances in the statistical and econometric methods of financial economics. The interactions between statistical biases and data mining in stock return studies should be a fruitful area for future research. Research on predictability in asset markets is likely to continue, and remain both useful and controversial for some time.

Acknowledgements

The author acknowledges financial support from the Collins Chair in Finance at Boston College.

References

Amihud, Y. and Horvitz, C. (2004). Predictive regressions: a reduced-bias estimation method. *Journal of Financial and Quantitative Analysis*, 39:813–842.

Ang, A., Chen, J. and Xing, Y. (2001). Downside risk and the momentum effect. NBER Working Paper No. 8643.

Avramov, D. and Chordia, T. (2006). Asset pricing models and financial market anomalies. *Review of Financial Studies*, 19:1001–1040.

Baker, M. and Wurgler, J. (2000). The equity share in new issues and aggregate stock returns. *Journal of Finance*, 55:2219–2257.

Barberis, N., Shleifer, A. and Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics*, 49:307–343.

Berk, J. (1995). A critique of size-related anomalies. *Review of Financial Studies*, 8:275–286.

Boudoukh, J. and Richardson, M. (1994). The statistics of long-horizon regressions revisited. *Mathematical Finance*, 4:103–120.

Boudoukh, J., Richardson, M. and Whitelaw, R. (1994). A tale of three schools: insights on the autocorrelations of short-horizon returns. *Review of Financial Studies*, 7:539–573.

Boudoukh, J., Richardson, M. and Whitelaw, R. (2004). The myth of long-horizon predictability. NBER Working Paper No. 11 841.

Bouman, S. and Jacobsen, B. (2002). The Halloween indicator: sell in May and go away: another puzzle. *American Economic Review*, 92:1618–1635.
Breen, W., Glosten, L. and Jagannathan, R. (1993). On the relation between the volatility and the expected value of the nominal excess returns on stocks. *Journal of Finance*, XLVII:1779–1801.
Campbell, J. Y. (1993). Intertemporal asset pricing without consumption data. *American Economic Review*, 83:487–512.
Campbell, J. Y. and Shiller, R. (1988). The dividend price ratio and expectations of future dividends and discount factors. *Review of Financial Studies*, 1:195–228.
Campbell, J. Y and Thompson, S. (2005). Predicting the equity premium out of sample: can anything beat the historical average? NBER Working Paper No. 11 468.
Campbell, J. Y. and Viceira, L. (2002). Strategic Asset Allocation: Portfolio Choice for Long-term Investors. New York, NY: Oxford University Press.
Chordia, T. and Shivakumar, L. (2002). Momentum, business cycle and time-varying expected returns. *Journal of Finance*, 57:985–1019.
Chui, A., Titman, S. and Wei, K. C. J. (2000). Momentum, ownership structure and financial crises: an analysis of Asian stock markets. Working Paper, University of Texas at Austin.
Cochrane, J. H. (2005). *Asset Pricing*. Princeton, NJ: Princeton University Press.
Cochrane, J. H. (2006). The dog that did not bark: a defense of return predictability. NBER Working Paper No. 12 026.
Cochrane, J. Y. (1988). How big is the random walk in GNP? *Journal of Political Economy*, 96: 893–920.
Conrad, J. and Kaul, G. (1988). Time-variation in expected returns. *Journal of Business*, 61:409–425.
Conrad, J. and Kaul, G. (1989). Mean reversion in short-horizon expected returns. *Review of Financial Studies*, 2:225–240.
Daniel, K., Hirshleifer, D. and Subrahmanyam, A. (1998). Investor psychology and security market under and over-reaction. *Journal of Finance*, 53:1839–1885.
DeBondt, W. and Thaler, R. (1985). Does the stock market overreact? *Journal of Finance*, 40:793–805.
Dybvig, P. H. and Ingersoll, J. (1982). Mean variance theory in complete markets. *Journal of Business*, 55:233–252.
Engle, R. (2004). Risk and volatility: econometric modelling and financial market practice. *American Economic Review*, 94:405–420.
Fama, E. F. (1970). Efficient capital markets: a review of theory and empirical work. *Journal of Finance*, 25:383–417.
Fama, E. and French, K. (1988). Permanent and temporary components of stock prices. *Journal of Political Economy*, 96:246–273.
Fama, E. and French, K. (1989). Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics*, 25:23–49.
Fama, E. F. and Schwert, G. W. (1977). Asset returns and inflation. *Journal of Financial Economics*, 5:115–146.
Ferson, W. E. (1989). Changes in expected security returns, risk and the level of interest rates. *Journal of Finance*, 44:1191–1217.
Ferson, W. E. (1995). Theory and empirical testing of asset pricing models. In: R. Jarrow, V. Maksimovic and B. Ziemba (eds), *Finance, Handbooks in Operations Research and Management Science*. Amsterdam: Elsevier, pp. 145–200.
Ferson, W. E. and Harvey, C. R. (1991). The variation of economic risk premiums. *Journal of Political Economy*, 99:385–415.
Ferson, W. E. and Harvey, C. R. (1993). The risk and predictability of international equity returns. *Review of Financial Studies*, 6:527–566.
Ferson, W. E. and Harvey, C. R. (1999). Conditioning variables and cross-section of stock returns. *Journal of Finance*, 54:1325–1360.
Ferson, W. and Korajczyk, R. (1995). Do arbitrage pricing models explain the predictability of stock returns? *Journal of Business*, 68:309–349.
Ferson, W. E. and Schadt, R. W. (1996). Measuring fund strategy and performance in changing economic conditions. *Journal of Finance*, 51:425–462.
Ferson, W. E., Heuson, A. and Su, T. (2005). Weak and semi-strong form stock return predictability revisited. *Management Science*, 51:1582–1592.
Ferson, W. E., Sarkissian, S. and Simin, T. (2003). Spurious regressions in financial economics? NBER Working Paper No. W9143.
Fisher, L. (1966). Some new stock market indexes. *Journal of Business*, 39:191–225.

Fleming, J., Kirby, C. and Ostdiek, B. (2001). The economic value of volatility timing. *Journal of Finance*, 56:329–352.
Foster, D., Smith, T. and Whaley, R. (1997). Assessing goodness-of-fit of asset pricing models: the distribution of the maximal R-squared. *Journal of Finance*, 52:591–607.
French, K., Schwert, G. W. and Stambaugh, R. (1987). Expected stock returns and volatility. *Journal of Financial Economics*, 19:3–29.
Friedman, M. (1953). The case for flexible exchange rates. In: *Essays in Positive Economics*. Chicago, IL: University of Chicago Press, pp. 157–203.
Gordon, M. J. (1962). *The Investment, Financing and Valuation of the Corporation*. Homewood, IL: Irwin.
Goyal, A. and Santa Clara, P. (2003) Idiosyncratic risk matters. *Journal of Finance*, 43:975–1007.
Goyal, A. and Welch, I. (2003). Predicting the equity premium with dividend ratios. *Management Science*, 49:639–654.
Goyal, A. and Welch, I. (2004). A comprehensive look at the empirical performance of equity premium prediction. NBER Working Paper No. 10 483.
Granger, C. W. K. and Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, 2:111–120.
Greenwood, R. (2004). Aggregate corporate liquidity and stock returns. Working Paper, Harvard University.
Grinblatt, M. and Han, B. (2003). The disposition effect and momentum. NBER Working Paper No. 8734.
Grinblatt, M. and Moskowitz, T. (2003). Predicting price movements from past returns: the role of consistency in tax-loss selling. *Journal of Financial Economics*, 71:541–579.
Harvey, C. R. (1991). The world price of covariance risk. *Journal of Finance*, 46:111–157.
Harvey, C. R. and Siddique, A. (2001). Autoregressive conditional skewness. *Journal of Financial and Quantitative Analysis*, 34:465–487.
Haugen, R. and Lakonishok, J. (1988). *The Incredible January Effect*. Homewood, IL: Dow Jones-Irwin.
Hjalmarsson, E. (2006). Should we expect significant out-of-sample results when predicting stock returns? Working Paper, Federal Reserve Board International Finance Discussion Paper No. 855.
Hodrick, R. (1992). Dividend yields and expected stock returns: alternative preocedures for inference and measurement. *Review of Financial Studies*, 5:257–286.
Hong, H. and Stein, J. (1999). A unified theory of underreaction, momentum trading and overreaction in asset markets. *Journal of Finance*, 45:265–295.
Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *Journal of Finance*, 45:881–898.
Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: implications for stock market efficiency. *Journal of Finance*, 48:65–91.
Jegadeesh, N. and Titman, S. (2001). Profitability of momentum portfolios: an evaluation of alternative explanations. *Journal of Finance*, 56:699–720.
Kandel, S. and Stambaugh, R. (1990). Expectations and volatility of consumption and asset returns. *Review of Financial Studies*, 3:207–232.
Kandel, S. and Stambaugh, R. (1996). On the predictability of stock returns: an asset allocation perspective. *Journal of Finance*, 51:385–424.
Kaul, G. (1996). Predictable components in stock returns. In: G. S. Maddala and C. R. Rao (eds), *Handbook of Statistics 14*. Amsterdam: Elsevier, pp. 269–296.
Keim, D. and Stambaugh, R. (1986). Predicting returns in the stock and bond markets. *Journal of Financial Economics*, 17:357–390.
Kendal, M.G. (1954). A note on bias in the estimation of autocorrelation. *Biometrika*, 41:403–404.
Korajczyk, R. and Sadka, R. (2004). Are momentum profits robust to trading costs? *Journal of Finance*, 59:1039–1082.
Kothari, S. P. and Shanken, J. (1997). Book-to-market, dividend yield and expected market returns: a time-series analysis. *Journal of Financial Economics*, 44:169–203.
Lamont, O. (2000). Investment plans and stock returns. *Journal of Finance*, 40(6):2719–2745.
Lehmann, B. N. (1990). Fads, martingales and market efficiency. *Quarterly Journal of Economics*, 105:1–28.
Lei, Q. (2006). Pricing dynamics of dividend and nondividend payout yields: on the stock return predictability. Working Paper, Southern Methodist University, Dallas.
Lesmond, D., Schill, M. and Zhou, C. (2004). The illusory nature of momentum profits. *Journal of Financial Economics*, 71:349–380.
Lettau, M. and Ludvigson, S. (2001). Consumption, aggregate wealth and expected stock returns. *Journal of Finance*, 56:815–626.

Lo, A. and MacKinlay, A. C. (1988). Stock market prices do not follow random walks: evidence from a simple specification test. *Review of Financial Studies*, 1:41–66.
Lo, A. and MacKinlay, A. C. (1990). An econometric analysis of infrequent trading. *Journal of Econometrics*, 45:181–211.
Malkiel, B. (2004). Models of stock market predictability. *Journal of Financial Research*, 28:449–459.
McLean, D. (2006). Why do momentum and long-term reversals persist? An evaluation of mispricing explanations. Working Paper, University of Alberta.
Merton, R. C. (1973). An intertemporal capital asset pricing model. *Econometrica*, 41:867–887.
Merton, R. C. (1980). On estimating the expected return on the market: an exploratory investigation. *Journal of Financial Economics*, 8:323–361.
Muthaswamy, J. (1988). Asynchronous closing prices and spurious correlation in portfolio returns. Working Paper, University of Chicago.
Pesaran, M. H. and Timmermann, A. (1995). Predictability of stock returns: robustness and economic significance. *Journal of Finance*, 50:1201–1228.
Pontiff, J. and Schall, L. (1998). Book-to-market as a predictor of market returns. *Journal of Financial Economics*, 49:141–160.
Pontiff, J. and Woodgate, A. (2006). Share issuance and cross-sectional stock returns. Working Paper, Boston College.
Powell, J. G, Shi, J., Smith, T. and Whaley, R. (2006). Political regimes, business cycles, seasonalities and returns. Working paper, Vanderbilt University, Nashville.
Richardson, M. (1993). Temporary components of stock prices: a skeptic's view. *Journal of Business and Economic Statistics*, 11:199–207.
Roll, R. R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *Journal of Finance*, 39:1127–1140.
Rouwehorst, G. (1998). International momentum portfolios. *Journal of Finance*, 53:267–284.
Rozeff, P. (1984). Dividend yields are equity premiums. *Journal of Portfolio Management*, 1(1):68–75.
Samuelson, P. (1965). Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review*, 6(2):41–49.
Santa-Clara, P. and Valkanov, R. (2003). The presidential puzzle: political cycles and the stock market. *Journal of Finance*, 58:1841–1872.
Scholes, M. and Williams, J. (1977). Estimating beta from nonsynchronous data. *Journal of Financial Economics*, 5:309–327.
Schwert, G. W. (2003). Anomalies and market efficiency. In: G. M. Constantinides, M. Harris and R. M. Stulz (eds), *Handbook of the Economics of Finance*. Amsterdam: Elsevier, pp. 575–603.
Scruggs, J. (1998). Resolving the puzzling intertemporal relation between the market risk premium and conditional market variance: a 2-factor approach. *Journal of Finance*, 53:575–603.
Sharpe, W. F. (1964). Capital asset prices: a theory of market equilibrium under conditions of risk. *Journal of Finance*, 19:425–442.
Simin, T. (2006). The (poor) predictive performance of asset pricing models. *Journal of Financial and Quantitative Analysis*, forthcoming.
Solnik, B. (1993). The unconditional performance of international asset allocation strategies using conditioning information. *Journal of Empirical Finance*, 1:33–55.
Stambaugh, R. S. (1999). Predictive regressions. *Journal of Financial Economics*, 54:375–421.
White, H. (2000). A reality check for data snooping. *Econometrica*, 68:1097–1126.
Working, H. (1949). The investigation of economic expectations. *American Economic Review*, 39:150–166.
Yule, G. (1926). Why do we sometimes get nonsense correlations between time series? A study in sampling and the nature of time-series. *Journal of the Royal Statistical Society*, 89:1–64.
Zarowin, P. (1990). Size, seasonality and stock market overreaction. *Journal of Financial and Quantitative Analysis*, 25:113–125.

2 A step-by-step guide to the Black–Litterman model

Incorporating user-specified confidence levels

Thomas Idzorek

Abstract

The Black–Litterman model enables investors to combine their unique views regarding the performance of various assets with the market equilibrium in a manner that results in intuitive, diversified portfolios. This paper consolidates insights from the relatively few works on the model, and provides step-by-step instructions that enable the reader to implement this complex model. A new method for controlling the tilts and the final portfolio weights caused by views is introduced. The new method asserts that the magnitude of the tilts should be controlled by the user-specified confidence level, based on a 0% to 100% confidence level. This is an intuitive technique for specifying one of most abstract mathematical parameters of the Black–Litterman model.

Having attempted to decipher many of the articles about the Black–Litterman model, none of the relatively few articles provide enough step-by-step instructions for the average practitioner to derive the new vector of expected returns.[1] *This article touches on the intuition of the Black–Litterman model, consolidates insights contained in the various works on the Black–Litterman model, and focuses on the details of actually combining market equilibrium expected returns with 'investor views' to generate a new vector of expected returns. Finally, I make a new contribution to the model by presenting a method for controlling the magnitude of the tilts caused by the views that is based on an intuitive 0% to 100% confidence level, which should broaden the usability of the model beyond quantitative managers.*

2.1 Introduction

The Black–Litterman asset allocation model, created by Fischer Black and Robert Litterman, is a sophisticated portfolio construction method that overcomes the problem

[1]The one possible exception to this is Robert Litterman's book *Modern Investment Management: An Equilibrium Approach*, published in July 2003 (the initial draft of this paper was written in November 2001), although I believe most practitioners will find it difficult to tease out enough information to implement the model. Chapter 6 of Litterman (2003) details the calculation of global equilibrium expected returns, including currencies; Chapter 7 presents a thorough discussion of the Black–Litterman model; and Chapter 13 applies the Black–Litterman framework to optimum active risk budgeting.

of unintuitive, highly-concentrated portfolios, input-sensitivity, and estimation error maximization. These three related and well-documented problems with mean-variance optimization are the most likely reasons that more practitioners do not use the Markowitz paradigm, in which return is maximized for a given level of risk. The Black–Litterman model uses a Bayesian approach to combine the subjective views of an investor regarding the expected returns of one or more assets with the market equilibrium vector of expected returns (the prior distribution) to form a new, mixed estimate of expected returns. The resulting new vector of returns (the posterior distribution) leads to intuitive portfolios with sensible portfolio weights. Unfortunately, the building of the required inputs is complex and has not been thoroughly explained in the literature.

The Black–Litterman asset allocation model was introduced in Black and Litterman (1990), expanded in Black and Litterman (1991, 1992), and discussed in greater detail in Bevan and Winkelmann (1998), He and Litterman (1999) and Litterman (2003).[2] The Black–Litterman model combines the CAPM (see Sharpe, 1964), reverse optimization (see Sharpe, 1974), mixed estimation (see Theil, 1971, 1978), the universal hedge ratio/Black's global CAPM (see Black, 1989a, 1989b; Litterman, 2003), and mean-variance optimization (see Markowitz, 1952).

Section 2.2 illustrates the sensitivity of mean-variance optimization and how reverse optimization mitigates this problem, while the following section presents the Black–Litterman model and the process of building the required inputs. Section 2.4 develops an implied confidence framework for the views. This framework leads to a new, intuitive method for incorporating the level of confidence in investor views that helps investors control the magnitude of the tilts caused by views. Section 2.5 concludes the chapter.

2.2 Expected returns

The Black–Litterman model creates stable, mean-variance efficient portfolios, based on an investor's unique insights, which overcome the problem of input-sensitivity. According to Lee (2000), the Black–Litterman model also 'largely mitigates' the problem of estimation error-maximization (see Michaud, 1989) by spreading the errors throughout the vector of expected returns.

The most important input in mean-variance optimization is the vector of expected returns; however, Best and Grauer (1991) demonstrate that a small increase in the expected return of one of the portfolio's assets can force half of the assets from the portfolio. In a search for a reasonable starting point for expected returns, Black and Litterman (1992), He and Litterman (1999) and Litterman (2003) explore several alternative forecasts: historical returns, equal 'mean' returns for all assets, and risk-adjusted equal mean returns. They demonstrate that these alternative forecasts lead to extreme portfolios – when unconstrained, portfolios with large long and short positions; and, when subject to a long only constraint, portfolios that are concentrated in a relatively small number of assets.

[2]Other important works on the model include Lee (2000), Satchell and Scowcroft (2000) and, for the mathematically inclined, Christodoulakis (2002).

2.2.1 Reverse optimization

The Black–Litterman model uses 'equilibrium' returns as a neutral starting point. Equilibrium returns are the set of returns that clear the market. The equilibrium returns are derived using a reverse optimization method in which the vector of implied excess equilibrium returns is extracted from known information using equation (2.1):[3]

$$\Pi = \lambda \Sigma w_{mkt} \tag{2.1}$$

where Π is the implied excess equilibrium return vector ($N \times 1$ column vector), λ is the risk-aversion coefficient, Σ is the covariance matrix of excess returns ($N \times N$ matrix), and w_{mkt} is the market capitalization weight ($N \times 1$ column vector) of the assets.[4]

The risk-aversion coefficient (λ) characterizes the expected risk-return tradeoff. It is the rate at which an investor will forego expected return for less variance. In the reverse optimization process, the risk-aversion coefficient acts as a scaling factor for the reverse optimization estimate of excess returns; the weighted reverse optimized excess returns equal the specified market risk premium. More excess return per unit of risk (a larger lambda) increases the estimated excess returns.[5]

To illustrate the model, I present an eight asset example in addition to the general model. To keep the scope of the paper manageable, I avoid discussing currencies.[6]

Table 2.1 presents four estimates of expected excess return for the eight assets – US Bonds, International Bonds, US Large Growth, US Large Value, US Small Growth, US Small Value, International Developed Equity and International Emerging Equity. The first CAPM excess return vector in Table 2.1 is calculated relative to the UBS Global Securities Markets Index (GSMI), a global index and a good proxy for the world market portfolio. The second CAPM excess return vector is calculated relative to the market

[3] Many of the formulas in this paper require basic matrix algebra skills. A sample spreadsheet is available from the author. Readers unfamiliar with matrix algebra will be surprised at how easy it is to solve for an unknown vector using Excel's matrix functions (MMULT, TRANSPOSE and MINVERSE). For a primer on Excel matrix procedures, go to http://www.stanford.edu/~wfsharpe/mia/mat/mia_mat4.htm.

[4] Possible alternatives to market capitalization weights include a presumed efficient benchmark and float-adjusted capitalization weights.

[5] The *implied* risk-aversion coefficient (λ) for a portfolio can be estimated by dividing the expected excess return by the variance of the portfolio (Grinold and Kahn, 1999):

$$\lambda = \frac{E(r) - r_f}{\sigma^2}$$

where $E(r)$ is the expected market (or benchmark) total return, r_f is the risk-free rate, and $\sigma^2 = w_{mkt}^T \Sigma w_{mkt}$ is the variance of the market (or benchmark) excess returns.

[6] Those who are interested in currencies are referred to Black (1989a, 1989b), Black and Litterman (1991, 1992), Grinold (1996), Meese and Crownover (1999), Grinold and Meese (2000), and Litterman (2003).

Table 2.1 Expected excess return vectors

Asset class	Historical μ_{Hist}	CAPM GSMI μ_{GSMI}	CAPM portfolio μ_P	Implied equilibrium return vector Π
US Bonds	3.15%	0.02%	0.08%	0.08%
Int'l Bonds	1.75%	0.18%	0.67%	0.67%
US Large Growth	−6.39%	5.57%	6.41%	6.41%
US Large Value	−2.86%	3.39%	4.08%	4.08%
US Small Growth	−6.75%	6.59%	7.43%	7.43%
US Small Value	−0.54%	3.16%	3.70%	3.70%
Int'l Dev. Equity	−6.75%	3.92%	4.80%	4.80%
Int'l Emerg. Equity	−5.26%	5.60%	6.60%	6.60%
Weighted average	−1.97%	2.41%	3.00%	3.00%
Standard deviation	3.73%	2.28%	2.53%	2.53%
High	3.15%	6.59%	7.43%	7.43%
Low	−6.75%	0.02%	0.08%	0.08%

All four estimates are based on 60 months of excess returns over the risk-free rate. The two CAPM estimates are based on a risk premium of 3. Dividing the risk premium by the variance of the market (or benchmark) excess returns (σ^2) results in a risk-aversion coefficient (λ) of approximately 3.07.

capitalization-weighted portfolio using *implied betas*, and is identical to the implied equilibrium return vector (Π).[7]

The historical return vector has a larger standard deviation and range than the other vectors. The first CAPM return vector is quite similar to the implied equilibrium return vector (Π) (the correlation coefficient is 99.8%).

Rearranging equation (2.1) and substituting μ (representing any vector of excess return) for Π (representing the vector of implied excess equilibrium returns) leads to equation (2.2), the solution to the unconstrained maximization problem: $\max_w w'\mu - \lambda w'\Sigma w/2$.

$$w = (\lambda\Sigma)^{-1}\mu \tag{2.2}$$

If μ does not equal Π, w will not equal w_{mkt}.

[7]Literature on the Black–Litterman model often refers to the reverse optimized implied equilibrium return vector (Π) as the CAPM returns, which can be confusing. CAPM returns based on regression-based betas can be significantly different from CAPM returns based on *implied betas*. I use the procedure in Grinold and Kahn (1999) to calculate implied betas. Just as it is possible to use the market capitalization weights and the covariance matrix to infer the implied equilibrium return vector, one can extract the vector of implied betas. The implied betas are the betas of the N assets relative to the market capitalization-weighted portfolio. As might be expected, the market capitalization-weighted beta of the portfolio is 1.

$$\beta = \frac{\Sigma w_{mkt}}{w_{mkt}^T \Sigma w_{mkt}} = \frac{\Sigma w_{mkt}}{\sigma^2}$$

where β is the vector of implied betas, Σ is the covariance matrix of excess returns, w_{mkt} is the market capitalization weights, and $\sigma^2 = w_{mkt}^T \Sigma w_{mkt} = \frac{1}{\beta^T \Sigma^{-1} \beta}$ is the variance of the market (or benchmark) excess returns.

The vector of CAPM returns is the same as the vector of reverse optimized returns when the CAPM returns are based on implied betas relative to the market capitalization-weighted portfolio.

Table 2.2 Recommended portfolio weights

Asset class	Weight based on historical w_{Hist}	Weight based on CAPM GSMI w_{GSMI}	Weight based on implied equilibrium return vector Π	Market capitalization weight w_{mkt}
US Bonds	1144.32%	21.33%	19.34%	19.34%
Int'l Bonds	−104.59%	5.19%	26.13%	26.13%
US Large Growth	54.99%	10.80%	12.09%	12.09%
US Large Value	−5.29%	10.82%	12.09%	12.09%
US Small Growth	−60.52%	3.73%	1.34%	1.34%
US Small Value	81.47%	−0.49%	1.34%	1.34%
Int'l Dev. Equity	−104.36%	17.10%	24.18%	24.18%
Int'l Emerg. Equity	14.59%	2.14%	3.49%	3.49%
High	1144.32%	21.33%	26.13%	26.13%
Low	−104.59%	−0.49%	1.34%	1.34%

In Table 2.2, equation (2.2) is used to find the optimum weights for three portfolios based on the return vectors from Table 2.1. The market capitalization weights are presented in the final column of Table 2.2.

Not surprisingly, the historical return vector produces an extreme portfolio. Those not familiar with mean-variance optimization might expect two highly correlated return vectors to lead to similarly correlated vectors of portfolio holdings. Nevertheless, despite the similarity between the CAPM GSMI return vector and the implied equilibrium return vector (Π), the return vectors produce two rather distinct weight vectors (the correlation coefficient is 66%). Most of the weights of the CAPM GSMI-based portfolio are significantly different than the benchmark market capitalization-weighted portfolio, especially the allocation to international bonds. As would be expected (since the process of extracting the implied equilibrium returns using the market capitalization weights is reversed), the implied equilibrium return vector (Π) leads back to the market capitalization-weighted portfolio. In the absence of views that differ from the implied equilibrium return, investors should hold the market portfolio. The implied equilibrium return vector (Π) is the market-neutral starting point for the Black–Litterman model.

2.3 The Black–Litterman model

2.3.1 The Black–Litterman formula

Prior to advancing, it is important to introduce the Black–Litterman formula and provide a brief description of each of its elements. Throughout this article, K is used to represent the number of views and N is used to express the number of assets in the formula. The formula for the new combined return vector ($E[R]$) is

$$E[R] = \left[(\tau\Sigma)^{-1} + P'\Omega^{-1}P\right]^{-1}\left[(\tau\Sigma)^{-1}\Pi + P'\Omega^{-1}Q\right] \tag{2.3}$$

where $E[R]$ is the new (posterior) combined return vector ($N \times 1$ column vector), τ is a scalar, Σ is the covariance matrix of excess returns ($N \times N$ matrix), P is a matrix that

identifies the assets involved in the views ($K \times N$ matrix or $1 \times N$ row vector in the special case of 1 view), Ω is a diagonal covariance matrix of error terms from the expressed views representing the uncertainty in each view ($K \times K$ matrix), Π is the implied equilibrium return vector ($N \times 1$ column vector), and Q is the view vector ($K \times 1$ column vector).

2.3.2 Investor views

More often than not, investment managers have specific views regarding the expected return of some of the assets in a portfolio, which differ from the implied equilibrium return. The Black–Litterman model allows such views to be expressed in either absolute or relative terms. Below are three sample views expressed using the format of Black and Litterman (1990).

- View 1: International developed equity will have an absolute excess return of 5.25% (confidence of view = 25%)
- View 2: International bonds will outperform US Bonds by 25 basis points (confidence of view = 50%)
- View 3: US Large Growth and US Small Growth will outperform US Large Value and US Small Value by 2% (confidence of view = 65%).

View 1 is an example of an absolute view. From the final column of Table 2.1, the implied equilibrium return of International Developed Equity is 4.80%, which is 45 basis points lower than the view of 5.25%.

Views 2 and 3 represent relative views. Relative views more closely approximate the way investment managers feel about different assets. View 2 says that the return of International Bonds will be 0.25% greater than the return of US Bonds. In order to gauge whether View 2 will have a positive or negative effect on International Bonds relative to US Bonds, it is necessary to evaluate the respective implied equilibrium returns of the two assets in the view. From Table 2.1, the implied equilibrium returns for International Bonds and US Bonds are 0.67% and 0.08%, respectively, for a difference of 0.59%. The view of 0.25%, from View 2, is less than the 0.59% by which the return of International Bonds exceeds the return of US Bonds; thus, one would expect the model to tilt the portfolio away from International Bonds in favour of US Bonds. In general (and in the absence of constraints and additional views), if the view is less than the difference between the two implied equilibrium returns, the model tilts the portfolio toward the underperforming asset, as illustrated by View 2. Likewise, if the view is greater than the difference between the two implied equilibrium returns, the model tilts the portfolio toward the outperforming asset.

View 3 demonstrates a view involving multiple assets, and that the terms 'outperforming' and 'underperforming' are relative. The number of outperforming assets need not match the number of assets underperforming. The results of views that involve multiple assets with a range of different implied equilibrium returns can be less intuitive. The assets of the view form two separate mini-portfolios – a long portfolio and a short portfolio. The relative weighting of each nominally outperforming asset is proportional to that asset's market capitalization divided by the sum of the market capitalization of the other nominally outperforming assets of that particular view. Likewise, the relative weighting of each nominally underperforming asset is proportional to that asset's market capitalization

Table 2.3a View 3 – nominally 'outperforming' assets

Asset class	Market capitalization (billions)	Relative weight	Implied equilibrium return vector Π	Weighted excess return
US Large Growth	$5174	90.00%	6.41%	5.77%
US Small Growth	$575	10.00%	7.43%	0.74%
	$5749	100.00%	Total	6.52%

Table 2.3b View 3 – nominally 'underperforming' assets

Asset class	Market capitalization (billions)	Relative weight	Implied equilibrium return vector Π	Weighted excess return
US Large Value	$5174	90.00%	4.08%	3.67%
US Small Value	$575	10.00%	3.70%	0.37%
	$5749	100.00%	Total	4.04%

divided by the sum of the market capitalizations of the other nominally underperforming assets. The net long positions less the net short positions equal 0. The mini-portfolio that actually receives the positive view may not be the nominally outperforming asset(s) from the expressed view. In general, if the view is greater than the weighted average implied equilibrium return differential, the model will tend to overweight the 'outperforming' assets.

From View 3, the nominally 'outperforming' assets are US Large Growth and US Small Growth, and the nominally 'underperforming' assets are US Large Value and US Small Value. From Table 2.3a, the weighted average implied equilibrium return of the mini-portfolio formed from US Large Growth and US Small Growth is 6.52%; from Table 2.3b, the weighted average implied equilibrium return of the mini-portfolio formed from US Large Value and US Small Value is 4.04%. The weighted average implied equilibrium return differential is 2.47%.

Because View 3 states that US Large Growth and US Small Growth will outperform US Large Value and US Small Value by only 2% (a reduction from the current weighted average implied equilibrium differential of 2.47%), the view appears to actually represent a reduction in the performance of US Large Growth and US Small Growth relative to US Large Value and US Small Value. This point is illustrated later in the chapter, in the final column of Table 2.6, where the nominally outperforming assets of View 3 – US Large Growth and US Small Growth – receive reductions in their allocations, and the nominally underperforming assets – US Large Value and US Small Value – receive increases in their allocations.

2.3.3 Building the inputs

One of the more confusing aspects of the model is moving from the stated views to the inputs used in the Black–Litterman formula. First, the model does not require that

investors specify views on all assets. In the eight-asset example, the number of views (k) is 3; thus, the view vector (Q) is a 3×1 column vector. The uncertainty of the views results in a random, unknown, independent, normally-distributed error term vector (ε) with a mean of 0 and covariance matrix Ω. Thus, a view has the form $Q + \varepsilon$.

General case: Example:

$$Q+\varepsilon=\begin{bmatrix} Q_1 \\ \vdots \\ Q_k \end{bmatrix}+\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_k \end{bmatrix} \qquad Q+\varepsilon=\begin{bmatrix} 5.25 \\ 0.25 \\ 2 \end{bmatrix}+\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_k \end{bmatrix} \tag{2.4}$$

Except in the *hypothetical* case, in which a clairvoyant investor is 100% confident in the expressed view, the error term (ε) is a positive or negative value other than 0. The error term vector (ε) does not directly enter the Black–Litterman formula. However, the variance of each error term (ω), which is the absolute difference from the error term's (ε) expected value of 0, does enter the formula. The variances of the error terms (ω) form Ω, where Ω is a diagonal covariance matrix with 0s in all of the off-diagonal positions. The off-diagonal elements of Ω are 0s because the model assumes that the views are independent of one another. The variances of the error terms (ω) represent the uncertainty of the views. The larger the variance of the error term (ω), the greater the uncertainty of the view.

General case:

$$\Omega=\begin{bmatrix} \omega_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \omega_k \end{bmatrix} \tag{2.5}$$

Determining the *individual* variances of the error terms (ω) that constitute the diagonal elements of Ω is one of the most complicated aspects of the model. It is discussed in greater detail below, and is the subject of Section 2.4.

The expressed views in column vector Q are matched to specific assets by matrix P. Each expressed view results in a $1 \times N$ row vector. Thus, K views result in a $K \times N$ matrix. In the three-view example presented in Section 2.2, in which there are 8 assets, P is a 3×8 matrix.

General case: Example (based on Satchell and Scowcroft, 2000):

$$P=\begin{bmatrix} p_{1,1} & \cdots & p_{1,n} \\ \vdots & \ddots & \vdots \\ p_{k,1} & \cdots & p_{k,n} \end{bmatrix} \qquad P=\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & .5 & -.5 & .5 & -.5 & 0 & 0 \end{bmatrix} \tag{2.6}$$

The first row of matrix P represents View 1, the absolute view. View 1 only involves one asset: International Developed Equity. Sequentially, International Developed Equity is the seventh asset in this eight-asset example, which corresponds with the '1' in the seventh column of Row 1. View 2 and View 3 are represented by Row 2 and Row 3, respectively. In the case of relative views, each row sums to 0. In matrix P, the nominally

outperforming assets receive positive weightings, while the nominally underperforming assets receive negative weightings.

Methods for specifying the values of matrix P vary. Litterman (2003: 82) assigns a percentage value to the asset(s) in question. Satchell and Scowcroft (2000) use an equal weighting scheme, which is presented in Row 3 of equation (2.6). Under this system, the weightings are proportional to 1 divided by the number of respective assets outperforming or underperforming. View 3 has two nominally underperforming assets, each of which receives a $-.5$ weighting. View 3 also contains two nominally outperforming assets, each receiving a $+.5$ weighting. This weighting scheme ignores the market capitalization of the assets involved in the view. The market capitalizations of the US Large Growth and US Large Value asset classes are nine times the market capitalizations of US Small Growth and Small Value asset classes, yet the Satchell and Scowcroft method affects their respective weights equally, causing large changes in the two smaller asset classes. This method may result in undesired and unnecessary tracking error.

Contrasting with the Satchell and Scowcroft (2000) equal weighting scheme, I prefer to use a market capitalization weighting scheme. More specifically, the relative weighting of each individual asset is proportional to the asset's market capitalization divided by the total market capitalization of either the outperforming or underperforming assets of that particular view. From the third columns of Tables 2.3a and 2.3b, the relative market capitalization weights of the nominally outperforming assets are 0.9 for US Large Growth and 0.1 for US Small Growth, while the relative market capitalization weights of the nominally underperforming assets are $-.9$ for US Large Value and $-.1$ for US Small Value. These figures are used to create a new matrix P, which is used for all of the subsequent calculations.

Matrix P (Market capitalization method):

$$P = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & .9 & -.9 & .1 & -.1 & 0 & 0 \end{bmatrix} \tag{2.7}$$

Once matrix P is defined, the variance of each individual view portfolio can be calculated. The variance of an individual view portfolio is $p_k \Sigma p_k'$, where p_k is a single $1 \times N$ row vector from matrix P that corresponds to the kth view and Σ is the covariance matrix of excess returns. The variances of the individual view portfolios ($p_k \Sigma p_k'$) are presented in Table 2.4. The respective variance of each individual view portfolio is an important source of information regarding the certainty, or lack thereof, of the level of confidence that should be placed on a view. This information is used shortly to revisit the variances of the error terms (ω) that form the diagonal elements of Ω.

Table 2.4 Variance of the view portfolios

View	Formula	Variance
1	$p_1 \Sigma p_1'$	2.836%
2	$p_2 \Sigma p_2'$	0.563%
3	$p_3 \Sigma p_3'$	3.462%

Conceptually, the Black–Litterman model is a complex, weighted average of the implied equilibrium return vector (Π) and the view vector (Q), in which the relative weightings are a function of the scalar (τ) and the uncertainty of the views (Ω). Unfortunately, the scalar and the uncertainty in the views are the most abstract and difficult to specify parameters of the model. The greater the level of confidence (certainty) in the expressed views, the closer the new return vector will be to the views. If the investor is less confident in the expressed views, the new return vector should be closer to the implied equilibrium return vector (Π).

The scalar (τ) is more or less inversely proportional to the relative weight given to the implied equilibrium return vector (Π). Unfortunately, guidance in the literature for setting the scalar's value is scarce. Both Black and Litterman (1992) and Lee (2000) address this issue: since the uncertainty in the mean is less than the uncertainty in the return, the scalar (τ) is close to zero. One would expect the equilibrium returns to be less volatile than the historical returns.[8]

Lee, who has considerable experience working with a variant of the Black–Litterman model, typically sets the value of the scalar (τ) between 0.01 and 0.05, and then calibrates the model based on a target level of tracking error (Dr Wai Lee, personal communication). Conversely, Satchell and Scowcroft (2000) say the value of the scalar (τ) is often set to 1.[9] Finally, Blamont and Firoozy (2003) interpret $\tau\Sigma$ as the standard error of estimate of the implied equilibrium return vector (Π); thus, the scalar (τ) is approximately 1 divided by the number of observations.

In the absence of constraints, the Black–Litterman model only recommends a departure from an asset's market capitalization weight if it is the subject of a view. For assets that are the subject of a view, the magnitude of their departure from their market capitalization weight is controlled by the ratio of the scalar (τ) to the variance of the error term (ω) of the view in question. The variance of the error term (ω) of a view is inversely related to the investor's confidence in that particular view. Thus, a variance of the error term (ω) of 0 represents 100% confidence (complete certainty) in the view. The magnitude of the departure from the market capitalization weights is also affected by other views. Additional views lead to a different combined return vector ($E[R]$), which leads to a new vector of recommended weights.

The easiest way to calibrate the Black–Litterman model is to make an assumption about the value of the scalar (τ). He and Litterman (1999) calibrate the confidence of a view so that the ratio of ω/τ is equal to the variance of the view portfolio ($p_k \Sigma p_k'$). Assuming $\tau = 0.025$ and using the individual variances of the view portfolios ($p_k \Sigma p_k'$) from Table 2.4, the covariance matrix of the error term (Ω) has the following form:

[8] The intuitiveness of this is illustrated by examining View 2, a relative view involving two assets of equal size. View 2 states that $p_2 \cdot E[R] = Q_2 + \varepsilon_2$, where $Q_2 = E[R_{Int'l.Bonds}] - E[R_{USBonds}]$. View 2 is $N \sim (Q_2, \omega_2)$. In the absence of additional information, it can be assumed that the uncertainty of the view is proportional to the covariance matrix (Σ). However, since the view is describing the mean return differential rather than a single return differential, the uncertainty of the view should be considerably less than the uncertainty of a single return (or return differential) represented by the covariance matrix (Σ). Therefore, the investor's views are represented by a distribution with a mean of Q and a covariance structure $\tau\Sigma$.

[9] Satchell and Scowcroft (2000) include an advanced mathematical discussion of one method for establishing a conditional value for the scalar (τ).

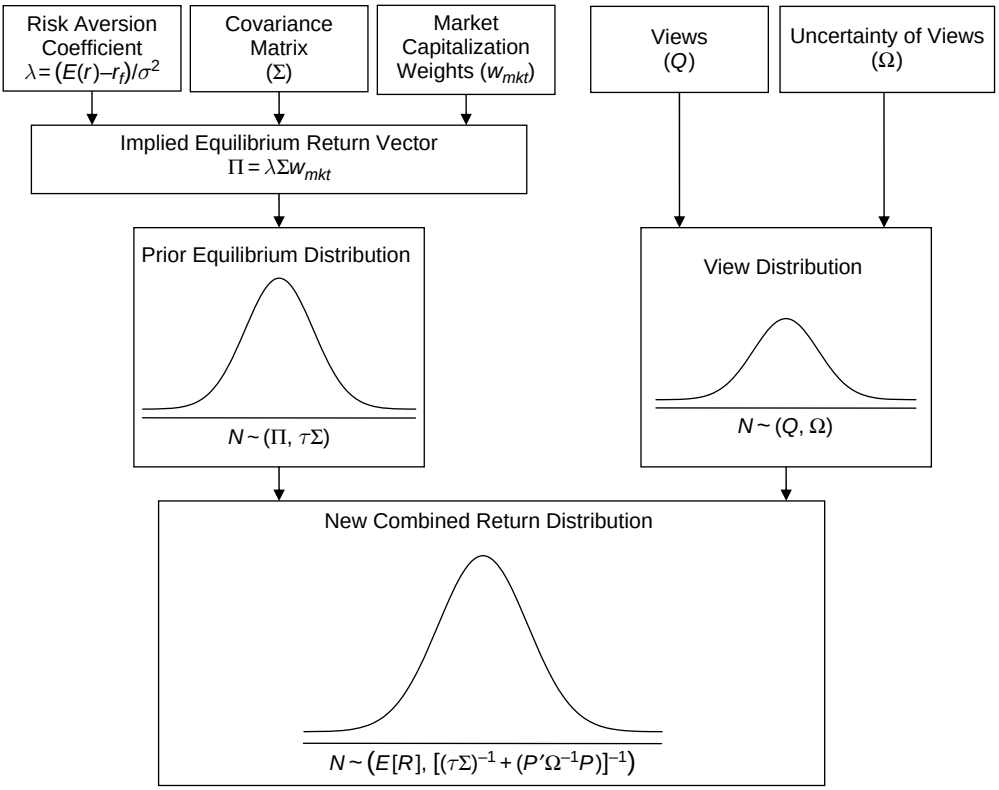
General case: Example:

$$\Omega = \begin{bmatrix} (p_1 \Sigma p_1') * \tau & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & (p_k \Sigma p_k') * \tau \end{bmatrix} \quad \Omega = \begin{bmatrix} 0.000709 & 0 & 0 \\ 0 & 0.000141 & 0 \\ 0 & 0 & 0.000866 \end{bmatrix} \tag{2.8}$$

When the covariance matrix of the error term (Ω) is calculated using this method, the actual value of the scalar (τ) becomes irrelevant because only the ratio ω/τ enters the model. For example, changing the assumed value of the scalar (τ) from 0.025 to 15 dramatically changes the value of the diagonal elements of Ω, but the new combined return vector ($E[R]$) is unaffected.

2.3.4 Calculating the new combined return vector

Having specified the scalar (τ) and the covariance matrix of the error term (Ω), all of the inputs are then entered into the Black–Litterman formula and the new combined return vector ($E[R]$) is derived. The process of combining the two sources of information is depicted in Figure 2.1. The new recommended weights ($\hat{w}$) are calculated by solving the



* The variance of the New Combined Return Distribution is derived in Satchell and Scowcroft (2000).

Figure 2.1 Deriving the New Combined Return Vector ($E[R]$).

unconstrained maximization problem, equation (2.2). The covariance matrix of historical excess returns (Σ) is presented in Table 2.5.

Even though the expressed views only directly involved seven of the eight asset classes, the individual returns of all the assets changed from their respective implied equilibrium returns (see column 4 of Table 2.6). A single view causes the return of every asset in the portfolio to change from its implied equilibrium return, since each individual return is linked to the other returns via the covariance matrix of excess returns (Σ).

The new weight vector ($\hat{w}$) in column 5 of Table 2.6 is based on the new combined return vector ($E[R]$). One of the strongest features of the Black–Litterman model is illustrated in the final column of Table 2.6. Only the weights of the seven assets for which views were expressed changed from their original market capitalization weights and the directions of the changes are intuitive.[10] No views were expressed on International Emerging Equity and its weights are unchanged.

From a macro perspective, the new portfolio can be viewed as the sum of two portfolios, where Portfolio 1 is the original market capitalization-weighted portfolio, and Portfolio 2 is a series of long and short positions based on the views. As discussed earlier, Portfolio 2 can be subdivided into mini-portfolios, each associated with a specific view. The relative views result in mini-portfolios with offsetting long and short positions that sum to 0. View 1, the absolute view, increases the weight of International Developed Equity without an offsetting position, resulting in portfolio weights that no longer sum to 1.

The intuitiveness of the Black–Litterman model is less apparent with added investment constraints, such as constraints on unity, risk, beta, and short selling. He and Litterman (1999) and Litterman (2003) suggest that, in the presence of constraints, the investor input the new combined return vector ($E[R]$) into a mean-variance optimizer.

2.3.5 Fine tuning the model

The Black–Litterman model can be fine tuned by studying the new combined return vector ($E[R]$), calculating the anticipated risk-return characteristics of the new portfolio, and then adjusting the scalar (τ) and the individual variances of the error term (ω) that form the diagonal elements of the covariance matrix of the error term (Ω).

Bevan and Winkelmann (1998) offer guidance in setting the weight given to the view vector (Q). After deriving an initial combined return vector ($E[R]$) and the subsequent optimum portfolio weights, they calculate the anticipated information ratio of the new portfolio. They recommend a maximum anticipated information ratio of 2.0. If the information ratio is above 2.0, decrease the weight given to the views (decrease the value of the scalar and leave the diagonal elements of Ω unchanged).

Table 2.7 compares the anticipated risk-return characteristics of the market capitalization-weighted portfolio with the Black–Litterman portfolio (the new weights

[10]The fact that only the weights of the assets that are subjects of views change from the original market capitalization weights is a criticism of the Black–Litterman model. Critics argue that the weight of assets that are highly (negatively or positively) correlated with the asset(s) of the view should change. I believe that the factors which lead to one's view would also lead to a view for the other highly (negatively or positively) correlated assets, and that it is better to make these views explicit.

Table 2.5 Covariance matrix of excess returns (Σ)

Asset class	US Bonds	Int'l Bonds	US Large Growth	US Large Value	US Small Growth	US Small Value	Int'l Dev. Equity	Int'l Emerg. Equity
US Bonds	0.001005	0.001328	−0.000579	−0.000675	0.000121	0.000128	−0.000445	−0.000437
Int'l Bonds	0.001328	0.007277	−0.001307	−0.000610	−0.002237	−0.000989	0.001442	−0.001535
US Large Growth	−0.000579	−0.001307	0.059852	0.027588	0.063497	0.023036	0.032967	0.048039
US Large Value	−0.000675	−0.000610	0.027588	0.029609	0.026572	0.021465	0.020697	0.029854
US Small Growth	0.000121	−0.002237	0.063497	0.026572	0.102488	0.042744	0.039943	0.065994
US Small Value	0.000128	−0.000989	0.023036	0.021465	0.042744	0.032056	0.019881	0.032235
Int'l Dev. Equity	−0.000445	0.001442	0.032967	0.020697	0.039943	0.019881	0.028355	0.035064
Int'l Emerg. Equity	−0.000437	−0.001535	0.048039	0.029854	0.065994	0.032235	0.035064	0.079958

Table 2.6 Return vectors and resulting portfolio weights

Asset class	New combined return vector $E[R]$	Implied equilibrium return vector Π	Difference $E[R]-\Pi$	New weight $\hat{w}$	Market capitalization weight w_{mkt}	Difference $\hat{w}-w_{mkt}$
US Bonds	0.07%	0.08%	−0.02%	29.88%	19.34%	10.54%
Int'l Bonds	0.50%	0.67%	−0.17%	15.59%	26.13%	−10.54%
US Large Growth	6.50%	6.41%	0.08%	9.35%	12.09%	−2.73%
US Large Value	4.32%	4.08%	0.24%	14.82%	12.09%	2.73%
US Small Growth	7.59%	7.43%	0.16%	1.04%	1.34%	−0.30%
US Small Value	3.94%	3.70%	0.23%	1.65%	1.34%	0.30%
Int'l Dev. Equity	4.93%	4.80%	0.13%	27.81%	24.18%	3.63%
Int'l Emerg. Equity	6.84%	6.60%	0.24%	3.49%	3.49%	0.00%
			Sum	103.63%	100.00%	3.63%

Table 2.7 Portfolio statistics

	Market capitalization-weighted portfolio w_{mkt}	Black–Litterman portfolio $\hat{w}$
Excess return	3.000%	3.101%
Variance	0.00979	0.01012
Standard deviation	9.893%	10.058%
Beta	1	1.01256
Residual return	–	0.063%
Residual risk	–	0.904%
Active return	–	0.101%
Active risk	–	0.913%
Sharpe ratio	0.3033	0.3083
Information ratio	–	0.0699

produced by the new combined return vector).[11] Overall, the views have very little effect on the expected risk-return characteristics of the new portfolio. However, both the Sharpe ratio and the information ratio increased slightly. The *ex ante* information ratio is well below the recommended maximum of 2.0.

Next, the results of the views should be evaluated to confirm that there are no unintended results. For example, investors confined to unity may want to remove absolute views, such as View 1.

Investors should evaluate their *ex post* information ratio for additional guidance when setting the weight on the various views. An investment manager who receives 'views' from a variety of analysts or sources could set the level of confidence of a particular view based in part on that particular analyst's information coefficient. According to Grinold and Kahn (1999), a manager's information coefficient is the correlation of forecasts with the actual results. This gives greater relative importance to the more skilful analysts.

Most of the examples in the literature, including the eight-asset example presented here, use a simple covariance matrix of historical returns. However, investors should use the best possible estimate of the covariance matrix of excess returns. Litterman and Winkelmann (1998) and Litterman (2003) outline the methods they prefer for estimating the covariance matrix of returns, as well as several alternative methods of estimation. Qian

[11]The data in Table 2.7 are based on the implied betas (see Note 7) derived from the covariance matrix of historical excess returns and the mean-variance data of the market capitalization-weighted benchmark portfolio. From Grinold and Kahn (1999):

$$\text{Residual return } \theta_P = E[R_P] - \beta_P * E[R_B]$$

$$\text{Residual risk } \omega_P = \sqrt{\sigma_P^2 - \beta_P^2 * \sigma_B^2}$$

$$\text{Active return } E[R_{PA}] = E[R_P] - E[R_B]$$

$$\text{Active risk } \Psi_P = \sqrt{\omega_P^2 + \beta_{PA}^2 * \sigma_B^2}$$

$$\text{Active portfolio beta } \beta_{PA} = (\beta_P - 1)$$

where $E[R_p]$ is the expected return of the portfolio, $E[R_B]$ is the expected return of the benchmark market capitalization-weighted portfolio based on the new combined expected return vector ($E[R]$), σ_B is the variance of the benchmark portfolio, and σ_P is the variance of the portfolio.

and Gorman (2001) extends the Black–Litterman model, enabling investors to express views on volatilities and correlations in order to derive a conditional estimate of the covariance matrix of returns. They assert that the conditional covariance matrix stabilizes the results of mean-variance optimization.

2.4 A new method for incorporating user-specified confidence levels

As the discussion above illustrates, Ω is the most abstract mathematical parameter of the Black–Litterman model. Unfortunately, according to Litterman (2003), how to specify the diagonal elements of Ω, representing the uncertainty of the views, is a common question without a 'universal answer.' Regarding Ω, Herold (2003) says that the major difficulty of the Black–Litterman model is that it forces the user to specify a probability density function for each view, which makes the Black–Litterman model only suitable for quantitative managers. This section presents a new method for determining the implied confidence levels in the views and how an implied confidence level framework can be coupled with an intuitive 0% to 100% user-specified confidence level in each view to determine the values of Ω, which simultaneously removes the difficulty of specifying a value for the scalar (τ).

2.4.1 Implied confidence levels

Earlier, the individual variances of the error term (ω) that form the diagonal elements of the covariance matrix of the error term (Ω) were based on the variances of the view portfolios ($p_k \Sigma p_k'$) multiplied by the scalar (τ). However, it is my opinion that there may be other sources of information in addition to the variance of the view portfolio ($p_k \Sigma p_k'$) that affect an investor's confidence in a view. When each view was stated, an intuitive level of confidence (0% to 100%) was assigned to each view. Presumably, additional factors can affect an investor's confidence in a view, such as the historical accuracy or *score* of the model, screen, or analyst that produced the view, as well as the difference between the view and the implied market equilibrium. These factors, and perhaps others, should be combined with the variance of the view portfolio ($p_k \Sigma p_k'$) to produce the best possible estimates of the confidence levels in the views. Doing so will enable the Black–Litterman model to maximize an investor's information.

Setting all of the diagonal elements of Ω equal to zero is equivalent to specifying 100% confidence in all of the K views. *Ceteris paribus*, doing so will produce the largest departure from the benchmark market capitalization weights for the assets named in the views. When 100% confidence is specified for all of the views, the Black–Litterman formula for the new combined return vector under 100% certainty ($E[R_{100\%}]$) is

$$E[R_{100\%}] = \Pi + \tau\Sigma P' \left(P\tau\Sigma P'\right)^{-1} \left(Q - P\Pi\right) \tag{2.9}$$

To distinguish the result of this formula from the first Black–Litterman formula (equation (2.3)), the subscript 100% is added. Substituting $E[R_{100\%}]$ for μ in equation (2.2) leads to $w_{100\%}$, the weight vector based on 100% confidence in the views. w_{mkt}, $\hat{w}$, and $w_{100\%}$ are illustrated in Figure 2.2.

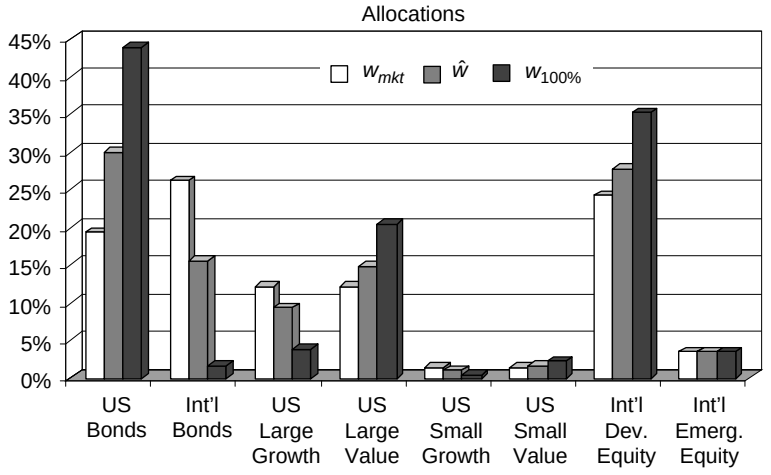


Figure 2.2 Portfolio allocations based on w_{mkt}, $\hat{w}$ and $w_{100\%}$.

When an asset is only named in one view, the vector of recommended portfolio weights based on 100% confidence ($w_{100\%}$) enables calculation of an intuitive 0% to 100% level of confidence for each view. In order to do so, the unconstrained maximization problem must be solved twice: once using $E[R]$ and once using $E[R_{100\%}]$. The new combined return vector ($E[R]$) based on the covariance matrix of the error term (Ω) leads to vector $\hat{w}$, while the new combined return vector ($E[R_{100\%}]$) based on 100% confidence leads to vector $w_{100\%}$. The departures of these new weight vectors from the vector of market capitalization weights (w_{mkt}) are $\hat{w} - w_{mkt}$ and $w_{100\%} - w_{mkt}$, respectively. It is then possible to determine the *implied* level of confidence in the views by dividing each weight difference ($\hat{w} - w_{mkt}$) by the corresponding maximum weight difference ($w_{100\%} - w_{mkt}$).

The implied level of confidence in a view, based on the scaled variance of the individual view portfolios derived in Table 2.4, is in the final column of Table 2.8. The implied confidence levels of View 1, View 2 and View 3 in the example are 32.94%, 43.06% and 33.02%, respectively. Only using the scaled variance of each individual view portfolio to determine the diagonal elements of Ω ignores the stated confidence levels of 25%, 50% and 65%.

Given the discrepancy between the stated confidence levels and the implied confidence levels, it is possible to experiment with different ωs, and recalculate the new combined return vector ($E[R]$) and the new set of recommended portfolio weights. I believe there is a better method.

2.4.2 The new method – an intuitive approach

I propose that the diagonal elements of Ω be derived in a manner that is based on the user-specified confidence levels and that results in portfolio tilts, which approximate $w_{100\%} - w_{mkt}$ multiplied by the user-specified confidence level (C).

$$Tilt_k \approx (w_{100\%} - w_{mkt}) * C_k \tag{2.10}$$

Table 2.8 Implied confidence level of views

Asset class	Market capitalization weights w_{mkt}	New weight $\hat{w}$	Difference $\hat{w} - w_{mkt}$	New weights (based on 100% confidence) $\hat{w}_{100\%}$	Difference $\hat{w}_{100\%} - w_{mkt}$	Implied confidence level $\frac{\hat{w} - w_{mkt}}{\hat{w}_{100\%} - w_{mkt}}$
US Bonds	19.34%	29.88%	10.54%	43.82%	24.48%	43.06%
Int'l Bonds	26.13%	15.59%	−10.54%	1.65%	−24.48%	43.06%
US Large Growth	12.09%	9.35%	−2.73%	3.81%	−8.28%	33.02%
US Large Value	12.09%	14.82%	2.73%	20.37%	8.28%	33.02%
US Small Growth	1.34%	1.04%	−0.30%	0.42%	−0.92%	33.02%
US Small Value	1.34%	1.65%	0.30%	2.26%	0.92%	33.02%
Int'l Dev. Equity	24.18%	27.81%	3.63%	35.21%	11.03%	32.94%
Int'l Emerg. Equity	3.49%	3.49%	–	3.49%	–	–

where $Tilt_k$ is the approximate tilt caused by the kth view ($N \times 1$ column vector), and C_k is the confidence in the kth view.

Furthermore, in the absence of other views, the approximate recommended weight vector resulting from the view is:

$$w_{k,\%} \approx w_{mkt} + Tilt_k \tag{2.11}$$

where $w_{k,\%}$ is the target weight vector based on the tilt caused by the kth view ($N \times 1$ column vector).

The steps of the procedure are as follows.

1. For each view (k), calculate the new combined return vector ($E[R_{k,100\%}]$) using the Black–Litterman formula under 100% certainty, treating each view as if it was the only view.

$$E[R_{k,100\%}] = \Pi + \tau\Sigma p_k' \left(p_k \tau\Sigma p_k'\right)^{-1} \left(Q_k - p_k\Pi\right) \tag{2.12}$$

   where $E[R_{100\%}]$ is the expected return vector based on 100% confidence in the kth view ($N \times 1$ column vector), p_k identifies the assets involved in the kth view ($1 \times N$ row vector), and Q_k is the kth View (1×1).
   Note: If the view in question is an absolute view and the view is specified as a total return rather than an excess return, subtract the risk-free rate from Q_k.
2. Calculate $w_{k,100\%}$, the weight vector based on 100% confidence in the kth view, using the unconstrained maximization formula.

$$w_{k,100\%} = (\lambda\Sigma)^{-1} E[R_{k,100\%}] \tag{2.13}$$

3. Calculate (pair-wise subtraction) the maximum departures from the market capitalization weights caused by 100% confidence in the kth view.

$$D_{k,100\%} = w_{k,100\%} - w_{mkt} \tag{2.14}$$

   where $D_{k,100\%}$ is the departure from market capitalization weight based on 100% confidence in kth view ($N \times 1$ column vector).
   Note: The asset classes of $w_{k,100\%}$ that are not part of the kth view retain their original weight leading to a value of 0 for the elements of $D_{k,100\%}$ that are not part of the kth view.
4. Multiply (pair-wise multiplication) the N elements of $D_{k,100\%}$ by the user-specified confidence (C_k) in the kth view to estimate the desired tilt caused by the kth view.

$$Tilt_k = D_{k,100\%} * C_k \tag{2.15}$$

   where $Tilt_k$ is the desired tilt (active weights) caused by the kth view ($N \times 1$ column vector), and C_k is an $N \times 1$ column vector where the assets that are part of the view receive the user-specified confidence level of the kth view and the assets that are not part of the view are set to 0.

5. Estimate (pair-wise addition) the target weight vector ($w_{k,\%}$) based on the tilt.

$$w_{k,\%} = w_{mkt} + Tilt_k \tag{2.16}$$

6. Find the value of ω_k (the kth diagonal element of Ω), representing the uncertainty in the kth view, that minimizes the sum of the squared differences between $w_{k,\%}$ and w_k.

$$\min \sum \left(w_{k,\%} - w_k\right)^2 \tag{2.17}$$

subject to $\omega_k > 0$ where

$$w_k = [\lambda\Sigma]^{-1}\left[(\tau\Sigma)^{-1} + p_k'\omega_k^{-1}p_k\right]^{-1}\left[(\tau\Sigma)^{-1}\Pi + p_k'\omega_k^{-1}Q_k\right] \tag{2.18}$$

Note: If the view in question is an absolute view and the view is specified as a total return rather than an excess return, subtract the risk-free rate from Q_k.[12]

7. Repeat steps 1–6 for the K views, build a $K \times K$ diagonal Ω matrix in which the diagonal elements of Ω are the ω_k values calculated in step 6, and solve for the new combined return vector ($E[R]$) using equation (2.3), which is reproduced here as equation (2.20).

$$E[R] = \left[(\tau\Sigma)^{-1} + P'\Omega^{-1}P\right]^{-1}\left[(\tau\Sigma)^{-1}\Pi + P'\Omega^{-1}Q\right] \tag{2.20}$$

Throughout this process, the value of the scalar (τ) is held constant and does not affect the new combined return vector ($E[R]$), which eliminates the difficulties associated with specifying it. Despite the relative complexities of the steps for specifying the diagonal elements of Ω, the key advantage of this new method is that it enables the user to determine the values of Ω based on an intuitive 0% to 100% confidence scale. Alternative methods for specifying the diagonal elements of Ω require that these abstract values be specified directly.[13] With this new method for specifying what was previously a very abstract mathematical parameter, the Black–Litterman model should be easier to use and more investors should be able to reap its benefits.

2.5 Conclusion

This chapter details the process of developing the inputs for the Black–Litterman model, which enables investors to combine their unique views with the implied equilibrium return

[12] Having just determined the weight vector associated with a specific view (w_k) in Step 6, it may be useful to calculate the active risk associated with the specific view in isolation.

$$\text{Active Risk created from } k\text{th view} = \sqrt{w_A^T \Sigma w_A} \tag{2.19}$$

where $w_A = w_k - w_{mkt}$ is the active portfolio weights, $w_k = [\lambda\Sigma]^{-1}[(\tau\Sigma)^{-1} + p_k'\omega_k^{-1}p_k]^{-1}[(\tau\Sigma)^{-1}\Pi + p_k'\omega_k^{-1}Q_k]$ is the weight vector of the portfolio based on the kth view and user-specified confidence level, and Σ is the covariance matrix of excess returns.

[13] Alternative approaches are explained in Fusai and Meucci (2003), Litterman (2003) and Zimmermann *et al.* (2002).

vector to form a new combined return vector. The new combined return vector leads to intuitive, well-diversified portfolios. The two parameters of the Black–Litterman model that control the relative importance placed on the equilibrium returns vs the view returns, the scalar (τ) and the uncertainty in the views (Ω), are very difficult to specify. The Black–Litterman formula with 100% certainty in the views enables determination of the implied confidence in a view. Using this implied confidence framework, a new method for controlling the tilts and the final portfolio weights caused by the views is introduced. The method asserts that the magnitude of the tilts should be controlled by the user-specified confidence level based on an intuitive 0% to 100% confidence level. Overall, the Black–Litterman model overcomes the most-often cited weaknesses of mean-variance optimization (unintuitive, highly concentrated portfolios, input-sensitivity, and estimation error maximization), helping users to realize the benefits of the Markowitz paradigm. Likewise, the proposed new method for incorporating user-specified confidence levels should increase the intuitiveness and the usability of the Black–Litterman model.

Acknowledgements

I am grateful to Robert Litterman, Wai Lee, Ravi Jagannathan, Aldo Iacono, and Marcus Wilhelm for helpful comments; to Steve Hardy, Campbell Harvey, Chip Castille and Barton Waring, who made this article possible and to the many others who provided me with helpful comments and assistance – especially my wife. Of course, all errors and omissions are my responsibility.

References

Best, M. J. and Grauer, R. R. (1991). On the sensitivity of mean-variance-efficient portfolios to changes in asset means: some analytical and computational results. *Review of Financial Studies*, 4(2):315–342.

Bevan, A. and Winkelmann, K. (1998). Using the Black–Litterman Global Asset Allocation Model: three years of practical experience. *Fixed Income Research*, Goldman, Sachs & Company, December.

Black, F. (1989a). Equilibrium exchange rate hedging. NBER Working Paper No. 2947.

Black, F. (1989b). Universal hedging: optimizing currency risk and reward in international equity portfolios. *Financial Analysts Journal*, 45:16–22.

Black, F. and Litterman, R. (1990). Asset allocation: combining investors views with market equilibrium. *Fixed Income Research*, Goldman, Sachs & Company, September.

Black, F. and Litterman, R. (1991). Global asset allocation with equities, bonds, and currencies. *Fixed Income Research*, Goldman, Sachs & Company, October.

Black, F. and Litterman, R. (1992). Global portfolio optimization. *Financial Analysts Journal*, 48(5):28–43.

Blamont, D. and Firoozy, N. (2003). Asset allocation model. *Global Markets Research: Fixed Income Research*, Deutsche Bank, July.

Christodoulakis, G. A. (2002). Bayesian optimal portfolio selection: the Black–Litterman approach. Unpublished paper (available online at http://www.staff.city.ac.uk/~gchrist/Teaching/QAP/optimalportfoliobl.pdf).

Fusai, G. and Meucci, A. (2003). Assessing views. *Risk*, 16(3):s18–21.

Grinold, R. C. (1996). Domestic grapes from imported wine. *Journal of Portfolio Management*, 26(Special issue):29–40.

Grinold, R. C. and Kahn, R. N. (1999). *Active Portfolio Management*, 2nd edn. New York, NY: McGraw-Hill.

Grinold, R. C. and Meese, R. (2000). The bias against international investing: strategic asset allocation and currency hedging. *Investment Insights*, Barclays Global Investors, August.

He, G. and Litterman, R. (1999). The intuition behind Black–Litterman model portfolios. *Investment Management Research*, Goldman, Sachs & Company, December.

Herold, U. (2003). Portfolio construction with qualitative forecasts. *Journal of Portfolio Management*, 29:61–72.
Lee, W. (2000). *Advanced Theory and Methodology of Tactical Asset Allocation*. New York, NY: John Wiley & Sons.
Litterman, R. and the Quantitative Resources Group, Goldman Sachs Asset Management. (2003). *Modern Investment Management: An Equilibrium Approach*. Princeton, NJ: John Wiley & Sons.
Litterman, R. and Winkelmann, K. (1998). Estimating covariance matrices. *Risk Management Series*, Goldman Sachs & Company, January.
Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance*, 7(1):77–91.
Meese, R. and Crownover, C. (1999). Optimal currency hedging. *Investment Insights*, Barclays Global Investors, April.
Michaud, R. O. (1989). The Markowitz optimization enigma: is optimized optimal? *Financial Analysts Journal*, 45:31–42.
Qian, E. and Gorman, S. (2001). Conditional distribution in portfolio theory. *Financial Analysts Journal*, 57(2):44–51.
Satchell, S. and Scowcroft, A. (2000). A demystification of the Black–Litterman model: managing quantitative and traditional construction. *Journal of Asset Management*, 1:138–150.
Sharpe, W. F. (1964). Capital asset prices: a theory of market equilibrium. *Journal of Finance*, 19(3):425–442.
Sharpe, W. F. (1974). Imputing expected security returns from portfolio composition. *Journal of Financial and Quantitative Analysis*, 9(3):463–472.
Theil, H. (1971). *Principles of Econometrics*. New York, NY: Wiley & Sons.
Theil, H. (1978). *Introduction to Econometrics*. Englewood Cliffs, NJ: Prentice-Hall.
Zimmermann, H., Drobetz, W. and Oertmann, P. (2002). *Global Asset Allocation: New Methods and Applications*. New York, NY: John Wiley & Sons.

3 A demystification of the Black–Litterman model: managing quantitative and traditional portfolio construction

Stephen Satchell and Alan Scowcroft

Abstract

The purpose of this paper is to present details of Bayesian portfolio construction procedures which have become known in the asset management industry as Black–Litterman models. We explain their construction, present some extensions and argue that these models are valuable tools for financial management.

3.1 Introduction

One of the major difficulties in financial management is trying to integrate quantitative and traditional management into a joint framework. Typically, traditional fund managers are resistant to quantitative management, as they feel that techniques of mean-variance analysis and related procedures do not capture effectively their value added. Quantitative managers often regard their judgemental colleagues as idiot savants. Senior management is rarely prepared to intervene when managers are successful and profitable, however they made their decisions. These disharmonies can make company-wide risk management and portfolio analysis non-operational and can have deleterious effects on company profitability and staff morale.

One model which has the potential to be used to integrate these diverse approaches is the Black–Litterman (BL) model (Black and Litterman, 1991, 1992). This is based on a Bayesian methodology which effectively updates currently held opinions with data to form new opinions. This framework allows the judgemental managers to give their views/forecasts, these views are added to the quantitative model and the final forecasts reflect a blend of both viewpoints. A lucid discussion of the model appears in Lee (1999).

Given the importance of this model, however, there appears to be no readable description of the mathematics underlying it. The purpose of this paper is to present such a description. In Sections 3.2 and 3.3 we describe the workings of the model and present some examples. In Section 3.4 we present an alternative formulation which takes into account prior beliefs on volatility. In Sections 3.2 and 3.3, particular attention is paid to the interesting issue of how to connect the subjective views of our managers into information usable in the model. This is not a trivial matter and lies at the heart of Bayesian

analysis. Indeed, Rev. Bayes had such misgivings about applying Bayes' theorem to real-world phenomena that he did not publish his paper (Bayes, 1763): it was presented to the Royal Society by his literary executor (Bernstein, 1996: 129–34).

3.2 Workings of the model

Before we present Bayes' theorem and its application by BL to asset management problems, we shall present our notation and basic concepts. We assume that there is an $(n \times 1)$ vector of asset returns $\mathbf{r}$; these are, typically, excess returns measured in the domestic currency and subtracting the domestic cash return which is not included in the vectors. With this convention the asset returns have a well-defined n-dimensional covariance matrix Σ; in particular, their covariance matrix is non-singular. If the returns for period t are denoted by $\mathbf{r}_t$, we shall write $E(\mathbf{r})$ to mean expected forecasted returns. This is shorthand for $E(\mathbf{r}_{t+1}|\Im_t)$, where $\Im_t$ refers to all information up to and including period t. A second related concept is the $(n \times 1)$ vector π representing equilibrium excess returns, either in terms of a theory such as the Capital Asset Pricing Model (CAPM) or in the sense of the prevailing supply of value-weighted assets. The latter interpretation corresponds to a global market portfolio demonetized in domestic currency.

Algebraically, assuming the validity of the CAPM,

$$\pi = \beta(\mu_m - r_f)$$

where μ_m is the return on the global market in domestic currency, r_f is the riskless (cash) domestic rate of return, β is an $(n \times 1)$ vector of asset betas, where

$$\beta = \mathrm{Cov}(\mathbf{r}, \mathbf{r}'\mathbf{w}_m)/\sigma_m^2$$

where $\mathbf{r}'\mathbf{w}_m$ is the return on the global market, $\mathbf{w}_m$ are the weights on the global market, determined by market values, and σ_m^2 is the variance of the rate of return on the world market.

If we let $\Sigma = \mathrm{Cov}(\mathbf{r}, \mathbf{r}')$ be the covariance matrix of the n asset classes, then

$$\pi = \delta\Sigma\mathbf{w}_m$$

where $\delta = (\mu_m - r_f)/\sigma_m^2$ is a positive constant. If returns were arithmetic with no reinvestment, δ would be invariant to time, since both numerator and denominator would be linear in time. However, if compounding is present, there may be some time effect.

In this chapter, we shall only consider (global) equity in our n assets. Extending the model to domestic and foreign equities and bonds presents few difficulties.

Considering Foreign Exchange (*FX*) as an additional asset class does present difficulties as we need to 'convert' currencies into a domestic value, which requires making assumptions about hedge ratios. Black (1990) proves that, in an international CAPM (ICAPM)[1]

[1]Here we use the acronym ICAPM to mean international CAPM. The standard usage for ICAPM is for intertemporal CAPM. Since the international CAPM is a particular application of Merton's intertemporal CAPM, this should cause no confusion.

under very stringent assumptions, all investors hedge the same proportion of overseas investment, and uses this result to justify a global or universal hedging factor which is the same for all investors facing all currencies. Adler and Prasad (1992) discuss Black's result and show how restrictive the result actually is.

It is natural to think of π as being the implied returns from the equilibrium model and, as the above discussion shows, these would depend upon our data and represent the input of the quantitative manager. How can we represent the views of the fund managers? To answer this question, consider Bayes' theorem. In the notation we have defined above, Bayes' theorem states that

$$\text{pdf}(E(\mathbf{r})|\pi) = \frac{\text{pdf}(\pi|E(\mathbf{r}))\text{pdf}(E(\mathbf{r}))}{\text{pdf}(\pi)}$$

where pdf(.) means probability density function. The above terms have the following interpretations:

- pdf($E(\mathbf{r})$) is the prior pdf that expresses the (prior) views of the fund manager/investor
- pdf(π) represents the marginal pdf of equilibrium returns. In the treatment that follows, it is not modelled. As we will demonstrate, it disappears in the constant of integration.
- pdf($\pi|E(\mathbf{r})$) is the conditional pdf of the data equilibrium return, given the forecasts held by the investor.

The result of the theorem pdf($E(\mathbf{r})|\pi$) is the 'combined' return or posterior forecast given the equilibrium information. It represents the forecasts of the manager/investor after updating for the information from the quantitative model.

The contribution of BL was to place this problem into a tractable form with a prior distribution that was both sensible and communicable to investors. Bayesian analysis has, historically, been weakened by difficulties in matching tractable mathematical distributions to individual's views.

We now review and extend BL's results. We make the following assumptions:

A1 pdf($E(\mathbf{r})$) is represented in the following way. The investor has a set of k beliefs represented as linear relationships. More formally, we know the $(k \times n)$ matrix $\mathbf{P}$ and a known $(k \times 1)$ vector $\mathbf{q}$. Let $\gamma = \mathbf{P}E(\mathbf{r})$ be a $(k \times 1)$ vector. It is assumed that $\mathbf{y} \sim N(\mathbf{q}, \Omega)$, where Ω is a $(k \times k)$ diagonal matrix with diagonal elements ω_{ii}. A larger ω_{ii} represents a larger degree of disbelief in the relationship represented by γ_i, $\omega_{ii} = 0$ represents absolute certainty, and, as a practical matter, we bound ω_{ii} above zero. The parameters $\mathbf{q}$ and Ω are called by Bayesian *hyperparameters*; they parameterize the prior pdf and are known to the investor.

A2 pdf($\pi|E(\mathbf{r})$) is assumed to be $N(E(\mathbf{r}), \tau\Sigma)$ where Σ is the covariance matrix of excess returns and τ is a (known) scaling factor often set to 1. This assumption means that the equilibrium excess returns conditional upon the individual's forecasts equals the individual's forecast on average. This may not hold in practice as the authors have met many practitioners who have exhibited the most alarming biases relative to the market view. The conditioning needs to be understood in the sense that, if all

individuals hold this view and invest in a CAPM-type world, then π represents the equilibrium returns conditional upon the individuals' common beliefs.[2]

Given A1 and A2, it is a straightforward result to show the following theorem.

Theorem 1: The pdf of $E(\mathbf{r})$ given π is given by

$$\begin{aligned}\text{pdf}(E(\mathbf{r})|\pi) \sim N([(\tau\Sigma)^{-1} + \mathbf{P}'\Omega^{-1}\mathbf{P}]^{-1} \\ [(\tau\Sigma)^{-1}\pi + \mathbf{P}'\Omega^{-1}\mathbf{q}], \\ [(\tau\Sigma)^{-1} + \mathbf{P}'\Omega^{-1}\mathbf{P}]^{-1}]\end{aligned}$$

Proof: See Appendix.

We emphasize that Theorem 1 is a result known to Bayesian econometricians and to BL, although they did not report the variance formula in the papers. Also, our interpretation of what is prior and what is sample information may differ from BL.

It should be clear from the previous analysis that neither A1 nor A2 are essential for the model to be used. Most priors used in finance, however, tend to convey little information about the investors' beliefs. Various alternatives such as a diffuse prior (see Harvey and Zhou, 1990: Equation 6; or Klein and Bawa, 1976: Equation 3) or the more detailed priors presented in Hamilton (1994: Chapter 12) cannot be easily understood in behavioural terms. In Bayesian terms, the prior chosen by BL is called the natural conjugate prior.

Extensions could be considered for volatilities as well. The natural *equilibrium* value for volatility is the Black–Scholes (BS) model, so that if option data were available, the prior on volatility could be updated by the *observed* implied volatility. Unfortunately, the *pdf* of implied volatility would depend on the nature of the stochastic volatility ignored by the BS formula, and there appears to be no simple way forward. An alternative would be to formulate a prior on τ. Although we have no obvious data to update our beliefs, a solution similar to Proposition 12.3 in Hamilton (1994) could be attempted. We present details in the Section 3.4.

3.3 Examples

In this section, we consider various examples which illustrate the methodology.

3.3.1 Example 1

In this example, we consider the case where a sterling-based investor believes that the Swiss equity market will outperform the German by 0.5% per annum. All returns are measured in sterling and are unhedged. This is a modest target, and is intended to emphasize that the forecast represents a new *equilibrium* and not short-term outperformance.

[2]This rather loose interpretation can be tightened; see Hiemstra (1997) for a construction of a CAPM model based on heterogeneous expectations by investors.

In the notation of the second section, we have one belief, $k = 1$ in A1. Using the universe of 11 European equity markets listed in Table 3.2, $\mathbf{P}$ is a (1×11) vector of the form

$$\mathbf{P} = [1, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0]$$

and $\mathbf{q} = 0.5\%$. Table 3.1 lists the parameters used to compute the conditional forecast.

The computed values $E(\mathbf{r}|\pi)$ are shown in Table 3.2. In addition to the prior view of the relative performance of Swiss and German markets, larger changes from the implied view for other markets are associated with low covariance with the Swiss market. In Table 3.3, we report certain key parameters associated with our portfolio construction.

We now consider the impact of the conditional forecast in an optimization problem, where the objective is a simple mean-variance utility function. The risk-aversion parameter

Table 3.1 Bayesian parameters

Parameter	Value	Symbol
Delta	5.00	δ
Tau	1.00	τ
View	0.05	q
Confidence	0.05	ω

Table 3.2 Forecast results

Market	Bench weight	Swiss Cov×100	π	$E(r\|\pi)$	Difference
Switzerland	0.0982	0.1884	5.34	5.53	+0.19
Germany	0.1511	0.0724	6.46	6.31	–0.15
Denmark	0.0136	0.0766	5.31	5.29	–0.02
Spain	0.0409	0.0666	8.07	7.99	–0.08
Finland	0.0125	0.0666	10.69	10.55	–0.14
France	0.1234	0.1016	7.93	7.89	–0.03
Italy	0.0568	0.0061	8.06	7.88	–0.18
Netherlands	0.0870	0.0826	5.64	5.62	–0.03
Norway	0.0103	0.0979	8.43	8.40	–0.03
Sweden	0.0474	0.0776	7.71	7.67	–0.04
UK	0.3588	0.0784	6.33	6.33	–0.00

Table 3.3 Optimization parameters

Parameter	Value
Risk aversion λ	2.5
Tracking error limit	2.5
Portfolio beta	1.0

Table 3.4 Optimization results

Market	Beta	Benchmark weight (%)	Solution weight (%)	Difference
Switzerland	0.80	9.82	12.19	+2.37
Germany	0.97	15.11	12.81	−2.30
Denmark	0.80	1.36	1.22	−0.14
Spain	1.20	4.09	4.27	+0.18
Finland	1.57	1.25	1.37	+0.11
France	1.18	12.34	12.61	+0.27
Italy	1.20	5.68	5.63	−0.05
Netherlands	0.85	8.70	8.04	−0.66
Norway	1.25	1.03	1.10	+0.07
Sweden	1.15	4.74	4.77	+0.02
UK	0.95	35.88	36.00	+0.12

has been set with reference to delta. The beta of the portfolio and the sum of the weights are constrained to unity. The results presented in Table 3.4, show, as expected, a switch from the German to the Swiss market. Some large differences in forecasts, Italy for example, are translated into small changes in the portfolio weights as the optimizer takes into account the benchmark weight, the asset beta and the impact of covariances.

In both this and the following example, currency holdings were free to vary between zero and minus the market weight (i.e. from unhedged to fully hedged). The assumed benchmark holding of currency is zero for all markets. The solution weights for currencies, which are not shown in the table, are all negligible.

The Sharpe ratio for the solution is 0.16 with a tracking error[3] of 0.39; the portfolio is beta constrained to 1.0.

3.3.2 Example 2

In this second example, we consider the case where a US dollar-based investor believes that six *hard currency* markets will outperform nine other European markets, on average, by 1.5% per annum. This could be interpreted as a possible EMU scenario. As in Example 1, this still represents one view and **P** is now a (1×15) vector equal to

$$[1/6 \ldots 1/6 - 1/9 \ldots - 1/9]$$

The values for the other parameters are as shown in Table 3.5. Note that delta (δ) has now been set at 3 to ensure that the level of the conditional forecast accords with historical experience.

The conditional forecast is shown in Table 3.6. The difference between the implied and conditional forecasts is broadly in line with the imposed view, with the exception that the forecast for Ireland actually goes down while Switzerland increases slightly.

[3]The tracking error or active risk of a portfolio is conventionally defined as the annualized standard deviation of portfolio active return (i.e. the excess return attributable to holding portfolio weights different from the benchmark weights).

Table 3.5 Bayesian parameters

Parameter	Value	Symbol
Delta	3.000	δ
Tau	1.000	τ
View	0.015	q
Confidence	0.025	ω

Table 3.6 Forecast results

	Bench weight	$\boldsymbol{\pi}$	$E(\mathbf{r}\|\boldsymbol{\pi})$	Difference
'Hard' markets				
Austria	0.0060	14.84	15.05	+0.21
Belgium	0.0244	13.75	13.83	+0.08
France	0.1181	14.86	14.98	+0.12
Germany	0.1446	13.57	13.60	+0.04
Netherlands	0.0832	12.33	12.38	+0.06
Ireland	0.0076	11.18	11.03	−0.15
'Soft' markets				
Denmark	0.0130	12.24	11.92	−0.32
Finland	0.0120	18.83	17.58	−1.24
Italy	0.0543	16.62	15.42	−1.20
Norway	0.0098	15.55	15.04	−0.51
Portugal	0.0049	11.84	11.67	−0.17
Spain	0.0392	13.63	13.03	−0.60
Sweden	0.0454	13.04	12.28	−0.76
Switzerland	0.0940	13.27	13.29	+0.02
UK	0.3433	12.74	12.68	−0.06

To consider the impact of the conditional forecast, we solve a simple optimization problem, where, as in Example 1, the asset weights are constrained to be positive and sum to unity. The asset beta is constrained to unity and currency weights are free to vary between unhedged and fully hedged for each market. The tracking error is bounded at 2.5 (see Table 3.7).

The optimization results are shown in Table 3.8 and, not surprisingly, show a positive tilt in favour of the *strong currency* markets. Interestingly, even though the optimizer was free to hold currency up to a fully hedged position, all the solution weights for currencies

Table 3.7 Optimization parameters

Parameter	Value
Risk aversion λ	1.5
Tracking error limit	2.5
Portfolio beta	1.0

Table 3.8 Optimization results

Market	Beta	Benchmark weight (%)	Solution weight (%)	Difference
Austria	1.09	0.60	3.27	+2.67
Belgium	1.02	2.44	4.82	+2.38
France	1.10	11.81	15.90	+4.09
Germany	1.00	14.46	18.63	+4.17
Netherlands	0.92	8.32	9.28	+0.96
Ireland	0.84	0.76	3.39	+2.63
Denmark	0.91	1.30	0.00	−1.30
Finland	1.37	1.20	0.00	−1.20
Italy	1.22	5.43	2.66	−2.77
Norway	1.14	0.98	0.00	−0.98
Portugal	0.88	0.49	0.00	−0.49
Spain	1.01	3.92	0.98	−2.94
Sweden	0.97	4.54	1.82	−2.72
Switzerland	0.98	9.40	6.79	−2.61
UK	0.95	34.33	32.46	−1.88

are zero. The Sharpe ratio for the solution is 0.18 with a tracking error of 1.8. The portfolio beta is constrained to unity.

Overall, we feel that the examples justify our confidence in the approach. Care needs to be taken interpreting the conditional forecast, however, since it is the product of the prior view *and* the data model. In these examples, the data model has been taken to be the implied excess returns generated by a mean-variance optimization problem. Even though such excess returns can be counter-intuitive, as in the case of Ireland in Example 2, they may be understood as the extent to which the neutral forecast has to change to reflect properly the views held by the investor. When these excess returns are subsequently fed back into the optimization process, the investor's optimal weights will reflect the prior view.

It is this usage of implied excess returns in the data model which also helps to address one of the principal reservations many practitioners have with respect to the use of optimizers in portfolio construction, namely their extreme sensitivity to changes in forecasts. Raw forecast alphas are inevitably volatile and, if used as optimizer inputs, give rise to completely unacceptable revisions to portfolio weights. By combining neutral model forecasts with the investor's views, the Bayesian formulation produces robust inputs for the optimizer.

3.4 Alternative formulations

In this section we present two alternative formulations of the BL model, the first of which takes into account prior beliefs about overall volatility. To do this, we make the following adjustment. We shall assume that τ is now unknown and stochastic so that

A3

$$\text{Pdf}(\boldsymbol{\pi}|E(\mathbf{r}), \tau) \sim N(E(\mathbf{r}),\ \tau\Sigma).$$

Furthermore,

$$\text{pdf}(E(\mathbf{r})|\tau) \sim N(\mathbf{q}, \tau\Omega),$$

A4 The marginal (prior) pdf of $\omega = 1/\tau$ is given by the following,[4]

$$\text{pdf}(\omega) = \frac{(\lambda/2)^{m/2}\omega^{(m/2)-1}\exp\left(-\frac{\lambda\omega}{2}\right)}{\Gamma(m/2)}, \qquad 0 < \omega < \infty$$

This pdf has two hyperparameters m and λ, and we assume it is independent of pdf($\boldsymbol{\pi}$).

Remark 1: Here we treat τ as a fundamental parameter that measures the overall dispersion of $\boldsymbol{\pi}$ about $E(\mathbf{r})$. Considering pdf($E(\mathbf{r})|\tau$), we define the elements of $\boldsymbol{\Omega}$ relative to τ so that $\omega_{ii} = 1$ reflects a degree of disbelief equal in scale to the dispersion measure of $\boldsymbol{\pi}$ about $E(\mathbf{r})$, a value $\omega_{ii} > 1$ implies greater disbelief than before and an increase in τ not only moves the dispersion of equilibrium expected returns about the forecasts but also increases the overall degree of disbelief in the forecasts.

Remark 2: The prior pdf of $\omega = 1/\tau$ is a scale gamma, where ω is often called the precision. It follows that $E(\omega) = m\lambda$ and $\text{Var}(\omega) = 2m\lambda^2$. This means that for fixed $E(\omega)$ as $m \to \infty$, $\text{Var}(\omega) \to 0$ and hence is a more reliable prior.

We are now in a position to state our new result which is, again, a standard result in the Bayesian literature. For a similar result, see Hamilton (1994: Proposition 12.3).

Theorem 2: If we assume A3 and A4, then

$$\text{pdf}(E(\mathbf{r})|\boldsymbol{\pi}) \propto$$
$$[m + (E(\mathbf{r}) - \theta)'\lambda^{\star}V(E(\mathbf{r}) - \theta)]^{-(m+n)/2}$$

which is a multivariate t distribution. The vector $\boldsymbol{\theta}$ is the term $E(\mathbf{r}|\boldsymbol{\pi})$ given in Theorem 1, the matrix V is the $\text{Var}(\mathbf{r}|\boldsymbol{\pi})$ given in Theorem 1 while

$$\lambda^{\star} = \frac{m}{\lambda + \mathbf{A} - \mathbf{C}'\mathbf{H}^{-1}\mathbf{C}}$$

where **A, C, H** are defined in the proof of Theorem 1.

[4] $\Gamma(.)$ is the gamma function, $\Gamma(n) = \int_0^{\infty} x^{n-1}\exp(-x)\mathrm{d}x$.

Proof: See Appendix.

An immediate corollary of Theorem 2 is the following.

Corollary 1: pdf$(\omega|E(\mathbf{r}), \pi)$ is a scale gamma with 'degrees of freedom' $m+n$ and scale factor $\mathbf{G}+\lambda$, where $\mathbf{G} = (\pi - E(\mathbf{r}))'\Sigma^{-1}(\pi - E(\mathbf{r})) + (\mathbf{P}E(\mathbf{r}) - \mathbf{q})'\boldsymbol{\Omega}^{-1}(\mathbf{P}E(\mathbf{r}) - \mathbf{q})$.

Proof: See Appendix.

The consequence of Corollary 2.1 is that we can now compute

$$E(\omega|E(\mathbf{r}), \pi) = (m+n)(\mathbf{G}+\lambda)$$

and

$$\text{Var}(\omega|E(\mathbf{r}), \pi) = 2(m+n)(\mathbf{G}+\lambda)^2$$

The increase in precision can be computed as

$$\begin{aligned} &E(\omega|E(\mathbf{r}), \pi) - E(w) \\ &\quad = (m+n)(\mathbf{G}+\lambda) - m\lambda \\ &\quad = m\mathbf{G} + n(\mathbf{G}+\lambda) \end{aligned}$$

It is interesting to note that, although our expected returns now have a multivariate t distribution, such a returns distribution is consistent with mean-variance analysis and the CAPM. (This is proved in Klein and Bawa, 1976.) Thus, our extended analysis leaves us with a mean vector and a covariance matrix which, up to a scale factor, are the same as before. What we gain is that probability computations will now involve the use of the t distribution. This will give the same probabilities as the normal for large m, but for small m will put more weights in the tails of our forecast distribution. Thus, we can manipulate this feature to give extra diagnostics to capture uncertainties about our forecasts.

We do not present numerical calculations for this model, as the nature of the prior is too complex to capture the beliefs of a typically non-mathematical fund manager. However, in our experience, fund managers are able to provide a range of scenarios for expected returns and associate probabilities with these scenarios. We shall explore such a model next, this being the second 'extension' of the BL model referred to earlier.

A5 The prior pdf for $E(\mathbf{r})$ is of the form $\mathbf{P}E(\mathbf{r}) = q_i,\ i = 1, \ldots, m$. Each (vector value q_i has prior probability p_i, where $\Sigma_{i=1}^{m} p_i = 1$. $\boldsymbol{P}$ and $E(\mathbf{r})$ have the same definition as before.

If we combine A5 with A3, it is straightforward to compute

$$p_i^\star = \text{prob}(\mathbf{P}E(\mathbf{r}) = q_i|\pi); \sum_{i=1}^{m} p_i^\star = 1$$

$$p_i^\star = \frac{\text{pdf}(\pi|\mathbf{P}E(\mathbf{r}) = q_i)\text{prob}(E(\mathbf{r}) = q_i)}{\text{pdf}(\pi)}$$

and

$$\text{pdf}(\pi) = \sum_{i=1}^{m} \text{pdf}(\pi|\boldsymbol{P}E(\boldsymbol{r}) = q_i)$$
$$\text{prob}(E(\mathbf{r}) = q_i)$$

$$\text{pdf}(\pi|\boldsymbol{P}E(\boldsymbol{r}) = q_i) = \varphi_{\text{i}}$$
$$= \left(\frac{1}{2\pi}\right)^{k/2} \frac{1}{\det(\boldsymbol{P}\Sigma\boldsymbol{P}')} \exp\left(\frac{-(\boldsymbol{P}\pi - q_i)'(\boldsymbol{P}\Sigma\boldsymbol{P}')^{-1}(\boldsymbol{P}\pi - q_i)}{2\tau}\right) \tag{3.1}$$

Combining the above, we deduce that $p_i^\star$ the posterior probability of scenario i becomes

$$p_i^\star = \frac{p_i\varphi_i}{\sum_{i=1}^{m} p_i\varphi_i} \tag{3.2}$$

Equation (3.2) gives us an updating rule on the prior probabilities which allows us to rescale our p_i by value of the likelihood function with expected returns evaluated at q_i normalized so that the sum of the weights is one. Thus, if the equilibrium return $\mathbf{p}$ satisfied the π condition $\mathbf{P}\pi = q_i$, φ_i would reach its maximum value. We note that since the term in front of the exponential in Equation (3.1) is common for all φ_i, we can simplify $p_i^\star$ to be

$$\frac{p_i \exp\left(\frac{-(\boldsymbol{P}\pi - q_i)'(\boldsymbol{P}\Sigma\boldsymbol{P}')^{-1}(\boldsymbol{P}\pi - q_i)}{2\tau}\right)}{\sum_{i=1}^{m} p_i \exp\left(\frac{-(\boldsymbol{P}\pi - q_i)'(\boldsymbol{P}\Sigma\boldsymbol{P}')^{-1}(\boldsymbol{P}\pi - q_i)}{2\tau}\right)} \tag{3.3}$$

Our new weights take a maximum value of 1 and a minimum value of 0. Table 3.9 provides an illustration of the calculations based on ten scenarios for the EMU example given in the third section. Column one shows the assumed outperformance of the strong currency markets for each scenario. For simplicity, we have assumed that the manager believes each scenario to be equally likely; $p_i = 0.1$. The calculated posterior probabilities $p_i^\star$ show clearly that substantial outperformance is much less likely given the historic covariances between these markets. In this example, each scenario is associated with only one view. If the scenario contained many views, the posterior probability would still relate to the entire scenario and not an individual view.

In practical terms, the judgemental fund manager can use the posterior probability $p_i^\star$ as a consistency check of the prior belief associated with scenario i expressed as probability p_i. If the scenario seems unlikely when tested against the data using equation (3.3) the confidence numbers ω_{ii} defined in A1 can be revised upwards accordingly. Equation (3.3) can therefore be regarded as a useful adjunct to Theorem 1 by helping the rational manager formulate the inputs required in a Bayesian manner. As observed by no less an authority than Harry Markowitz, 'the rational investor is a *Bayesian*' (Markowitz, 1987: 57, italics in original).

Table 3.9 Posterior probabilities

Scenario outperformance (%)	Prior probability (%)	Posterior probability (%)
0.5	10.00	10.07
1.0	10.00	10.06
1.5	10.00	10.05
2.0	10.00	10.04
2.5	10.00	10.02
3.0	10.00	10.00
3.5	10.00	9.98
4.0	10.00	9.96
4.5	10.00	9.93
5.0	10.00	9.90

3.5 Conclusion

We have presented several examples of Bayesian asset allocation portfolio construction models and showed how they combine judgemental and quantitative views. It is our belief that these models are potentially of considerable importance in the management of the investment process in modern financial institutions where both viewpoints are represented. We present an exposition of these models so that readers should be able to apply these methods themselves. We also present several extensions.

Appendix

Proof of Theorem 1

Using Bayes' theorem and Assumptions A1 and A2, we see that

$$\text{pdf}(E(\mathbf{r})|\pi) = \frac{k\ \exp(-\frac{1}{2\tau}(\pi - E(\mathbf{r}))'\Sigma^{-1}(\pi - E(\mathbf{r})) - \frac{1}{2}(\mathbf{P}E(\mathbf{r}) - \mathbf{q})'\Omega^{-1}(\mathbf{P}E(\mathbf{r}) - \mathbf{q}))}{\text{pdf}(\pi)}$$

where k is an appropriate constant.

We next simplify the quadratic term in the exponent.

$$\begin{aligned}
&E(\mathbf{r})'(\tau\Sigma)^{-1}E(\mathbf{r}) - 2\pi(\tau\Sigma)^{-1}E(\mathbf{r}) + \pi'(\tau\Sigma)^{-1}\pi + E(\mathbf{r})'\mathbf{P}'\boldsymbol{\Omega}^{-1}\mathbf{P}E(\mathbf{r}) \\
&\quad - 2\mathbf{q}'\boldsymbol{\Omega}^{-1}\mathbf{P}E(\mathbf{r}) + \mathbf{q}'\boldsymbol{\Omega}^{-1}\mathbf{q} \\
&\quad = E(\mathbf{r})'((\tau\Sigma)^{-1} + \mathbf{P}'\boldsymbol{\Omega}\mathbf{P})((\tau\Sigma)^{-1} + \mathbf{P}'\boldsymbol{\Omega}^{-1}\mathbf{P})^{-1}((\tau\Sigma)^{-1} + \mathbf{P}'\boldsymbol{\Omega}^{-1}\mathbf{P})E(\mathbf{r}) - 2(\pi'(\tau\Sigma)^{-1} \\
&\qquad + \mathbf{q}'\boldsymbol{\Omega}^{-1}\mathbf{P})((\tau\Sigma)^{-1} + \mathbf{P}'\boldsymbol{\Omega}^{-1}\mathbf{P})^{-1}((\tau\Sigma)^{-1} + \mathbf{P}'\boldsymbol{\Omega}^{-1}E(\mathbf{r})) + \mathbf{q}'\boldsymbol{\Omega}^{-1}\mathbf{q} + \pi'(\tau\Sigma)^{-1}\pi
\end{aligned}$$

Let

$$\mathbf{C} = (\tau\Sigma)^{-1}\pi + \mathbf{P}'\boldsymbol{\Omega}^{-1}\mathbf{q}$$

$$\mathbf{H} = (\tau\Sigma)^{-1} + \mathbf{P}'\boldsymbol{\Omega}^{-1}\mathbf{P}, \text{ we note that } \mathbf{H} \text{ is symmetrical so } \mathbf{H} = \mathbf{H}'$$

$$\mathbf{A} = \mathbf{q}'\boldsymbol{\Omega}^{-1}\mathbf{q} + \pi'(\tau\Sigma)^{-1}\pi$$

We can re-write the exponent as equal to

$$E(\mathbf{r})'\mathbf{H}'\mathbf{H}^{-1}\mathbf{H}E(\mathbf{r}) - 2\mathbf{C}'\mathbf{H}^{-1}\mathbf{H}E(\mathbf{r}) + \mathbf{A}$$
$$= (\mathbf{H}E(\mathbf{r}) - \mathbf{C})'\mathbf{H}^{-1}(\mathbf{H}E(\mathbf{r}) - \mathbf{C}) + \mathbf{A} - \mathbf{C}'\mathbf{H}^{-1}\mathbf{C}$$
$$= (E(\mathbf{r}) - \mathbf{H}^{-1}\mathbf{C})'\mathbf{H}(E(\mathbf{r}) - \mathbf{H}^{-1}\mathbf{C}) + \mathbf{A} - \mathbf{C}'\mathbf{H}^{-1}\mathbf{C}$$

In terms of $E(\mathbf{r})$, terms such as $\mathbf{A} - \mathbf{C}'\mathbf{H}\mathbf{C}$ disappear into the constant of integration. Thus,

$$\text{pdf}(E(\mathbf{r})|\pi) \propto \exp\left(-\frac{1}{2}(E(\mathbf{r}) - \mathbf{H}^{-1}\mathbf{C})'\mathbf{H}(E(\mathbf{r}) - \mathbf{H}^{-1}\mathbf{C})\right) \tag{A.1}$$

$$\text{so that } E(\mathbf{r})|\pi \text{ has mean} = \mathbf{H}^{-1}\ \mathbf{C} \tag{A.2}$$

$$= [(\tau\Sigma)^{-1} + \mathbf{P}'\boldsymbol{\Omega}^{-1}\mathbf{P}]^{-1}\ [(\tau\Sigma)^{-1}\pi + \mathbf{P}'\boldsymbol{\Omega}^{-1}\mathbf{q}] \tag{A.3}$$

$$\text{and} \qquad \text{Var}(\mathbf{r}|\pi) = [(\tau\Sigma)^{-1} + \mathbf{P}'\boldsymbol{\Omega}^{-1}\mathbf{P}]^{-1} \tag{A.4}$$

Proof of Theorem 2

First

$$\text{pdf}(E(\mathbf{r}), w|\pi) = \frac{\text{pdf}(\pi|E(\mathbf{r}), w)\text{pdf}(E(\mathbf{r})|w)\text{pdf}(w)}{\text{pdf}(\pi)} \tag{A.5}$$

From Assumption A3, we can write

$$\text{pdf}(\pi|E(\mathbf{r}), w)\text{pdf}(E(\mathbf{r})|w) = kw^{n/2}\exp - \left(\frac{w\mathbf{G}}{2}\right) \tag{A.6}$$

where $\mathbf{G} = (\pi - \text{E}(\mathbf{r}))'\Sigma^{-1}(\pi - E(\mathbf{r})) + (\mathbf{P}E(\mathbf{r}) - \mathbf{q})'\boldsymbol{\Omega}^{-1}(PE(\mathbf{r}) - \mathbf{q})$

If we now use Assumption 4 and Equation (A.6), we see that

$$\text{pdf}(E(\mathbf{r})), w|\pi) = \frac{k\exp\left(-\frac{w}{2}(\mathbf{G} + \boldsymbol{\lambda})\right)\left(\frac{\boldsymbol{\lambda}}{2}\right)^{m/2} w^{(m+n)/2-1}}{\Gamma\left(\frac{m}{2}\right)\text{pdf}(\pi)} \tag{A.7}$$

To compute $\text{pdf}(E(r)|\pi)$, we integrate out w.

Let

$$v = \frac{w}{2}(\mathbf{G}+\boldsymbol{\lambda}),\ w = \frac{2v}{(\mathbf{G}+\boldsymbol{\lambda})},\ \mathrm{d}w = \frac{2}{(\mathbf{G}+\boldsymbol{\lambda})}dv$$

$$\mathrm{pdf}(E(\mathbf{r})|\pi) = k'\left(\frac{\boldsymbol{\lambda}}{2}\right)^{m/2}\int_0^{\infty}\exp(-v)\left(\frac{2v}{\mathbf{G}+\boldsymbol{\lambda}}\right)^{(m+n)/2-1}\left(\frac{2}{\mathbf{G}+\boldsymbol{\lambda}}\right)\mathrm{d}v \tag{A.8}$$

then

$$= \frac{k'\left(\frac{\lambda}{2}\right)^{m/2}2^{(m+n)/2}\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)(\mathbf{G}+\lambda)^{(m+n)/2}} \tag{A.9}$$

The multivariate t is defined (see Zellner, 1971: 383, B20) for matrices $\boldsymbol{\theta}(l\times 1)$ and $\mathbf{V}(l\times 1)$ and positive constant v as

$$\mathrm{pdf}(\mathbf{x}|\boldsymbol{\theta},\mathbf{V},v,l) = \frac{v^{v/2}\Gamma((v+1)/2|\mathbf{V}|^{1/2}[v+(\mathbf{x}-\boldsymbol{\theta})'\mathbf{V}(\mathbf{x}-\theta)]^{(l+v)/2}}{\pi^{1/2}\Gamma(v/2)}$$

If we re-write $\mathbf{G}+\lambda$ in terms of **A, C** and **H** as defined in the proof of Theorem 1, we see that

$$\mathbf{G}+\lambda = (E(\mathbf{r})-\mathbf{H}^{-1}\mathbf{C})'\mathbf{H}(E(\mathbf{r})-\mathbf{H}^{-1}\mathbf{C})+\mathbf{A}-\mathbf{C}'\mathbf{H}^{-1}\mathbf{C}+\lambda$$

$$\propto (E(\mathbf{r})-\mathbf{H}^{-1}\mathbf{C})'\frac{m\mathbf{H}}{\lambda+\mathbf{A}-\mathbf{C}'\mathbf{H}^{-1}\mathbf{C}}(E(\mathbf{r})-\mathbf{H}^{-1}\mathbf{C})+m \tag{A.10}$$

This shows that $\mathrm{pdf}(E(\mathbf{r})|\pi)$ is multivariate t,

$$\boldsymbol{\theta} = \mathbf{H}^{-1}\mathbf{C} \text{ (as before)}$$

$$\mathrm{V} = \frac{m}{\boldsymbol{\lambda}+\mathrm{A}-\mathbf{C}'\mathbf{H}^{-1}\mathbf{C}}\mathbf{H},\ l = n \text{ and } v = m.$$

Proof of Corollary 2.1

Factorizing $\mathrm{pdf}(E(\mathbf{r}),w|\pi) = \mathrm{pdf}(w|E(\mathbf{r}),\pi)\mathrm{pdf}(E(\mathbf{r})|\pi)$ gives us

$$\mathrm{pdf}(w|E(\mathbf{r}),\pi) = \frac{\mathrm{pdf}(E(\mathbf{r}),w|\pi)}{\mathrm{pdf}(E(\mathbf{r})|\pi)}$$

thus

$$\mathrm{pdf}(w|E(\mathbf{r}),\pi) = \frac{\dfrac{k\exp\left(-\frac{w}{2}(\mathbf{G}+\lambda)\right)\left(\frac{\lambda}{2}\right)^{m/2}w^{(m+n)/2-1}}{\Gamma\left(\frac{m}{2}\right)\mathrm{pdf}(\pi)}}{\dfrac{k'\left(\frac{\lambda}{2}\right)^{m/2}2^{(m+n)/2}\Gamma\left(\frac{m+n}{2}\right)}{(\mathbf{G}+\boldsymbol{\lambda})^{(m+n)/2}\Gamma\left(\frac{m}{2}\right)\mathrm{pdf}(\pi)}} \tag{A.11}$$

using Equations (A.6) and A.7).

Simplifying,

$$\mathrm{pdf}(w|E(\mathbf{r}), \pi) = \frac{k'' \exp\left(-\frac{w}{2}(\mathbf{G}+\lambda)\right) w^{(m+n)/2-1}\mathbf{G}^{(m+n)/2}}{\Gamma\left(\frac{m+n}{2}\right)} \tag{A.12}$$

References

Adler, M. and Prasad, B. (1992). On universal currency hedges. *Journal of Financial and Quantitative Analysis*, 27(1):19–39.

Bayes, T. (1763). An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions*, Essay LII:370–418.

Bernstein, P. L. (1996). *Against the Gods*. Chichester: John Wiley.

Black, F. (1989). Universal hedging: optimal currency risk and reward in international equity portfolios. *Financial Analysts Journal*, July–Aug:16–22, reprinted in *Financial Analysts Journal*, Jan–Feb.:161–167.

Black, F. (1990). Equilibrium exchange rate hedging. *Journal of Finance*, 45(3):899–907.

Black, F. and Litterman, R. (1991). Global asset allocation with equities, bonds and currencies. Goldman Sachs and Co., October.

Black, F. and Litterman, R. (1992). Global portfolio optimization. *Financial Analysts Journal*, Sep–Oct:28–43.

Hamilton, J. (1994). *Time Series Analysis*. Princeton, NJ: Breedon University Press.

Harvey, C. R. and Zhou, G. (1990). Bayesian inference in asset pricing tests. *Journal of Financial Economics*, 26:221–254.

Hiemstra, C. (1997). *Addressing Nonlinear Dynamics in Asset Prices with a Behavioural Capital Asset Pricing Model*. University of Strathclyde: Accounting and Finance Department.

Klein, R. W. and Bawa, V. S. (1976). The effect of estimation risk on optimal portfolio choice. *Journal of Financial Economics*, 3:215–231.

Lee, W. (1999). Advanced theory and methodology of tactical asset allocation. Unpublished manuscript, available at http://faculty.fuqua.duke.edu/~charvey/Teaching/BA453_2000/lee.pdf

Markowitz, H. (1987). *Mean-Variance Analysis in Portfolio Choice and Capital Markets*. Oxford: Basil Blackwell.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York, NY: John Wiley.

4 Optimal portfolios from ordering information

Robert Almgren and Neil Chriss

Abstract

Modern portfolio theory produces an optimal portfolio from estimates of expected returns and a covariance matrix. We present a method for portfolio optimization based on replacing expected returns with *sorting criteria* – that is, with information about the order of the expected returns but not their values. We give a simple and economically rational definition of optimal portfolios that extends Markowitz' definition in a natural way; in particular, our construction allows full use of covariance information. We give efficient numerical algorithms for constructing optimal portfolios. This formulation is very general and is easily extended to more general cases, where assets are divided into multiple sectors or there are multiple sorting criteria available, and may be combined with transaction cost restrictions. Using both real and simulated data, we demonstrate dramatic improvement over simpler strategies.

4.1 Introduction

This chapter presents a framework for portfolio selection when an investor possesses information about the order of expected returns in the cross-section of stocks, but not the values of the expected returns. Even in the simplest case of a complete ordering, there has previously been no rational way to form an optimal portfolio that makes full use of covariance information; for the first time we give a complete solution to this important problem.

In general, ordering information may be any set of inequality beliefs about the expected returns, such as the order of expected returns across all stocks, sorts within sectors or other subdivisions, decile rankings, sorts with sign beliefs, multiple incompatible sorts, or incomplete information. We give a simple and general method of producing portfolios that are optimal with respect to this information.

Portfolio selection as introduced by Markowitz (1952) constructs portfolios by maximizing expected return subject to a set of constraints. His key contribution was the observation that an optimizing investor should want to invest only in *efficient* portfolios that deliver the maximum level of expected return for a given level of risk. In the absence of expected returns it is not clear how to generalize this approach. But ordering information has become increasingly important to the financial literature and the investment process: many researchers and practitioners have associated both firm characteristics and recent price history to expected returns in a manner that naturally gives rise to ordering

information (Banz, 1981; Fama and French, 1992, 1996; Campbell *et al.*, 1993; Daniel and Titman, 1997, 1998).

If only ordering information is given, there is no obvious objective function. One may construct an objective function in one of several ways, including estimating expected returns from the data that give rise to the ordering information. One can also develop *ad hoc* rules relating such characteristics to expected returns and proceed in the same fashion. Finally, one can simply propose rules for constructing portfolios from ordering information, such as buying the top decile of stocks and selling the bottom decile.

In this chapter we propose a portfolio selection procedure that assumes no information beyond the given ordering information and is based on a simple, economically rational set of assumptions. We do not rely on any expected return estimates or *ad hoc* rules to produce portfolios; our method moves naturally from ordering information to portfolio. We do not argue one way or the other whether one should or could obtain better results by estimating expected returns and then performing ordinary portfolio observation. We simply observe that there may be cases where either one cannot obtain enough data to make such estimates or one does not have sufficient confidence in the reliability of such estimates to warrant using this approach. Therefore our base assumption is that the investor is in possession of no information beyond the given ordering information.

Our approach starts with the observation that Markowitz portfolio selection may be viewed as a statement about investor preferences. All else being equal, an investor should prefer a portfolio with a higher expected return to one with a lower expected return. Efficient portfolios are simply portfolios that are *maximally preferable* among those with a fixed level of risk. This is a very slight change in the point of view offered by Markowitz, but yields a substantial generalization of Markowitz portfolio selection theory. In the present chapter we extend the notion of a portfolio preference to one based on ordering information instead of expected returns. Given such a preference relation we define an efficient portfolio as one that is again maximally preferable for a given level of risk exactly analogous to the Markowitz definition. Our challenge here is to provide a simple and economically rational definition of such a preference relation and then to demonstrate how to calculate efficient portfolios and that these portfolios have desirable properties.

We start with the observation that for any ordering information there is a set of expected returns that are consistent with this information. For example, if we start with three stocks S_1, S_2, S_3 and our information consists of the belief that $\rho_1 \geq \rho_2 \geq \rho_3$ where ρ_i is the expected return for S_i, then all triples of the form (r_1, r_2, r_3) where $r_1 \geq r_2 \geq r_3$ are consistent with the ordering. We write Q for the set of all expected returns that are consistent with the given ordering. We then state, loosely speaking, that given two portfolios w_1 and w_2 an investor should prefer w_1 to w_2 if w_1 has a higher expected return than w_2 for every consistent expected return – that is, for every $r \in Q$. The precise definition is given in Section 4.2 and involves an orthogonal decomposition of portfolio space based on the set of beliefs.

In Section 4.2 we study the above defined preference relation in detail and define an efficient portfolio to be one that is *maximally preferable* within a set of constraints. For example, if the covariance matrix of the stocks in the portfolio is given by V, then for a given level of variance σ^2 a portfolio w is efficient if

$$w^{\mathrm{T}} V w \leq \sigma^2$$

and there is no portfolio v such that v is preferable to w and $v^{\mathrm{T}} V v \leq \sigma^2$.

Our definition of efficient is a natural analog of the Markowitz notion of efficient; it reduces to exactly Markowitz' efficient set in the case of an information set consisting only of one expected return vector. We exploit the mathematical structure of the set Q to completely characterize the set of efficient portfolios. We also show how to construct efficient portfolios and how to extend these constructions to more general constraint sets including market-neutral portfolios with a given level of risk and portfolios constrained by transaction cost limits.

We note that the preference relation obtained from ordering information is not strong enough to identify a unique efficient portfolio for a given level of risk. For a fixed level of risk, Markowitz portfolio selection does identify a unique efficient portfolio; our process identifies an infinite set of portfolios which are all equally preferable. That said, the efficient set is extremely small relative to the entire constraint set. In fact, it is in the order of less than $1/n!$ the measure of the constraint set, where n is the number of stocks in the portfolio.

In Section 4.3 we refine the preference relation introduced in Section 4.2 to produce a unique efficient portfolio for each level of risk. This refinement starts with the observation that while we certainly prefer w_1 to w_2 when its expected return is greater for *all* consistent returns, we would in fact prefer it if its expected return were higher for a *greater fraction* of consistent returns. To make sense of this notion we have no choice but to introduce a probability measure on the set of expected returns which assesses the relative likelihood of different consistent returns being the true expected returns. Once we specify such a probability measure, we are able to produce our natural refinement of the preference relation.

We show that the preference relation thus obtained yields a unique efficient portfolio for each level of risk. Moreover, we show that when the probability distribution obeys certain very natural symmetry properties, the related preference relation is completely characterized by a certain linear function called the *centroid*. Specifically, this function is defined by the centre of mass, or centroid, of the set Q under the probability measure. We call the efficient portfolios with respect to the centroid preference relation *centroid optimal*.

A natural probability distribution on the set of consistent returns is one in which, roughly speaking, all expected returns are equally likely. This is most consistent with the notion that we have no information about expected returns beyond the ordering information. On the other hand, one might argue that, at least on a qualitative level, consistent returns that have very large or very small separation are relatively less likely than those which have moderate separation. For example, if the average return during the past twelve months among three stocks is 0.03, then while (0.0001, 0.0000001, 0.000000001) and (0.05, 0.03, 0.02) and (1000, 500, 200) all may be consistent, one might argue that the first and the last ought to be less likely than the second. Fortunately, we show that to a large extent such considerations do not affect the outcome of our analysis in the sense that for a large class of distributions, the centroid optimal portfolio is the same. In particular, one may construct distributions that assign different probabilities as one scales a particular consistent expected return. For example, if one has a certain probability on the interval $[0, \infty]$ then one can move that distribution to the ray $\lambda(0.05,\ 0.03,\ 0.02)$ for all $\lambda \geq 0$. If one extends this distribution so that it is the same along all such rays, then we show that the centroid optimal portfolio is independent of what distribution is chosen along each ray.

The rest of this chapter is organized as follows. In Section 4.2 we give a detailed discussion of preference relations and how they give rise to efficient portfolios. In Section 4.3 we derive the theory behind centroid optimal portfolios and explain how to calculate them. In Section 4.4 we study a variety of different types of ordering information, including sorts within sectors, sorts that arise from index outperform versus underform ratings, and multiple sorts that arise from multiple firm characteristics. In each case we demonstrate how this information fits into our framework and show how to calculate centroid optimal portfolios. In Section 4.5 we conduct two different studies. First, we demonstrate how to compute centroid optimal portfolios for *reversal strategies* – that is, strategies that involve ordering information derived from recent price action. We compare performance history on a cross-section of stocks for four different portfolio construction methodologies and show that the centroid optimal portfolios outperform by a substantial margin. Our second set of tests is based on simulated data. These tests are designed to examine the robustness of our methods relative to imperfect knowledge of the exact order of expected returns. We show that our method is highly robust to rather large levels of information degradation.

4.2 Efficient portfolios

This section studies portfolio selection from asset ordering information. We construct portfolios which, by analogy with modern portfolio theory (MPT), are *efficient* in the sense that they are maximally preferable to a rational investor for a fixed level of variance. The challenge in this section is to define a mathematically coherent economically sensible definition of *preferable* and *maximally preferable*. To do this requires appealing to the mathematics of convex optimization, and in particular the structure of convex cones. As a complete exposition of this subject is beyond the scope of this chapter, we refer readers to Boyd and Vandenberghe (2004) for details.

In MPT investors start with two sets of *probability beliefs*, concerning the first two moments of the return distributions of the stocks in their universe, and seek to find *efficient portfolios* that provide the maximum level of expected return for a given level of variance. In our setting these expected return beliefs are replaced by *ordering beliefs* whereby an investor has a set of beliefs about the order of the expected returns of a universe of assets.

We summarize the portfolio selection procedure in this section as follows. We start with a universe of stocks, a portfolio sort and a budget constraint set. A portfolio sort is information about an investor's belief as to the order of expected returns. This may be in the form of a complete or partial sort, and may possibly contain more than one sort. For example, an investor may have one sort arising from book to market value and another arising from market capitalization. The key example of a budget constraint is portfolio variance, but more generally a budget constraint is any bounded, convex subset of the space of all portfolios. Given this information, we use the portfolio sort to define a unique preference relation among portfolios whereby, roughly speaking, one portfolio is preferred to another if it has a higher expected return for *all* expected returns that are consistent with the sort. Given this preference relation and the budget constraint, we define a portfolio to be *efficient* if it is *maximally preferable* within the budget constraint set.

The aim of this section is to make precise the notion of an efficient portfolio and to show how we may calculate efficient portfolios. Along the way we show that our preference relation naturally extends the portfolio preference relation implicit in Markowitz' portfolio selection in which a rational investor prefers one portfolio to another if it has a higher expected return. The main difference that comes to light in building portfolios from sorts is that the efficient set for a fixed budget constraint (e.g. for a fixed level of portfolio variance) is no longer a unique portfolio but rather a bounded subset of the constraint set.

4.2.1 Modern portfolio theory

To motivate the development in this section, we briefly review modern portfolio theory, recast in terms that we can generalize to the case of portfolio selection from ordering information. We assume throughout that we have a list of assets in a fixed order and that vectors represent either expected returns for those assets or investments in those assets. Thus a vector $\rho = (\rho_1, \ldots, \rho_n)$ will represent expected return estimates for assets $1, \ldots, n$ respectively, and a vector $w = (w_1, \ldots, w_n)$ will refer to a portfolio of investments in assets $1, \ldots, n$ respectively.[1] Therefore the expected return of a portfolio w is w_ρ^{T}.

Markowitz introduced the notion of *efficient* portfolios: those that provide the maximum expected return subject to a given maximum level of variance. Optimizing investors should seek to invest only in efficient portfolios. Therefore, the Markowitz portfolio selection problem may be stated as the constrained optimization problem: *find the portfolio with the highest expected return for a given level of variance*. This problem can be easily recast in terms of preference relations.

Throughout this chapter, we will use the symbol $\succeq$ to denote preference between portfolios; $v \succeq w$ means that an optimizing investor prefers v to w. In MPT, for a given ρ, optimizing investors set their preferences on the basis of their expected returns subject to a risk constraint:

$$v \succeq w \text{ if and only if } v^{\mathrm{T}}\rho \geq w^{\mathrm{T}}\rho.$$

An investor seeking an efficient portfolio wants to invest in only those portfolios which are *maximally preferable* subject to having no more than a fixed level of variance.

This small change in point of view (from maximizing expected returns to finding most preferable portfolios) allows us to introduce the appropriate language for generalizing MPT to the case of portfolio selection from ordering information. To do this we introduce two new notions: *expected return cones* and *relevant* portfolio components.

Expected return cones

Let

$$\overline{Q} = \{\lambda\rho | \lambda \geq 0\} \tag{4.1}$$

[1] We take all vectors to be column vectors, and T denotes transpose, so the matrix product $w^{\mathrm{T}}r$ is equivalent to the standard Euclidean inner product $w \cdot r = \langle w, r \rangle$. This formality is useful when we extend our notation to multiple linear conditions.

be the smallest cone that contains ρ. We remind the reader that a subset Q of a vector space is a cone if for every $r \in Q$ and scalar $\lambda > 0$ we have $\lambda r \in Q$. Note that

$$v \succeq w \text{ if and only if } v^{\mathrm{T}} r \geq w^{\mathrm{T}} r \text{ for all } r \in \overline{Q}.$$

Thus in terms of setting investor preferences a specific expected return vector is interchangeable with the cone containing it. That is, it is not the magnitude of the expected return vector that determines investor preferences, but only its *direction*. In Section 4.2.2 we generalize this construction to the case of portfolio beliefs.

We also define the half-space whose inward normal is ρ:

$$\overline{Q} = \{r \in \mathbb{R}^n | \rho^{\mathrm{T}} r \leq 0\}$$

We will see later in this section that this half-space contains the same belief information as the vector ρ or the ray $\overline{Q}$ in terms of investor beliefs.

Relevant and irrelevant portfolio directions

We now define a decomposition of the space of portfolios into directions *relevant* and *irrelevant* to a given expected return vector ρ.

$$R^{\perp} = \{r \in \mathbb{R}^n | \rho^{\mathrm{T}} r = 0\} \tag{4.2}$$

This subspace is the complete collection of return vectors that are orthogonal to ρ. Let

$$R = \left(R^{\perp}\right)^{\perp} = \{w \in \mathbb{R}^n | w^{\mathrm{T}} r = 0 \text{ for all } r \in R^{\perp}\}$$

be the orthogonal subspace to $R^{\perp}$. This subspace contains ρ but is larger; in particular it contains *negative* multiples of the base vector ρ. We note that R and $R^{\perp}$ define a unique decomposition of the space of portfolios as follows. For a portfolio w we have that w may be written

$$w = w_0 + w_{\perp} \quad \text{with } w_0 \in R \text{ and } w_{\perp} \in R^{\perp}$$

R is the 'relevant' subspace defined by our beliefs, as expressed by the expected return vector ρ. It is relevant in the sense that its complementary component $w_{\perp}$ has zero expected return.

The portfolio preference relation defined above may be rewritten as

$$v \succeq w \text{ if and only if } v_0^{\mathrm{T}} r \geq w_0^{\mathrm{T}} r \text{ for all } r \in Q,$$

where w_0 and v_0 are the relevant components of w and v respectively. Note now we have two very different ways of expressing the same preference relation among portfolios. On the one hand we may compare portfolios v and w versus the returns in the cone $\overline{Q}$, while on the other hand we may compare the relevant parts of each portfolio versus all expected returns in the half-space Q. These two formulations are equivalent in terms of the preference relation they yield, but, as we shall see in Section 4.2.3, the half-space formulation may be generalized to the case of preference relations arising from portfolio sorts.

4.2.2 The set of returns consistent with a sort

In this section we tackle in more detail the notion of portfolio sort and formalize some notation. The key object of study is the set of expected returns which are consistent with a sort; intuitively, this is the set of all possible expected returns which *could be* the actual set of expected returns given an investor's beliefs.

We start with an investor who possesses a list of n stocks with expected return $r = (r_1, \ldots, r_n)$ and covariance matrix V. We assume that the investor does not know r but has m distinct *beliefs* about the relationship between the components of r, expressed by m different sets of inequalities. In this sense, each *belief* is a linear inequality relationship among the expected returns. As an example, a belief might be of the form $r_4 \geq r_8$ or $4r_2 + 2r_3 \geq r_4$. We restrict our attention to *homogeneous* linear relationships, with no constant term. Thus, for example, we do not allow beliefs of the form 'the average of r_1 and r_2 is at least 3% annual'.

Each belief may be expressed in a mathematically compact form as a linear combination of expected returns being greater than or equal to zero. For example,

$$4r_2 + 2r_3 - r_4 \geq 0$$

In this way we may place the coefficients of such inequalities into a column vector D_1 and write the inequality in the form $D_1^{\mathrm{T}} r \geq 0$. In the above example we would have

$$D_1 = (0, 4, 2, -1, 0, \ldots, 0)^{\mathrm{T}}$$

We can collect together all beliefs into m column vectors $D_1, \ldots, D_m$ containing the coefficients of the inequalities as above. We call each vector D_j a *belief vector*, and our aim is to look at these in their totality.

The total set of beliefs may succinctly be expressed as $Dr \geq 0$, where D is the $m \times n$ *belief matrix* whose rows are $D_1^{\mathrm{T}}, \ldots, D_m^{\mathrm{T}}$, and a vector is ≥ 0 if and only if each of its components is non-negative. We do not require that the belief vectors be independent, and we allow $m < n$, $m = n$ or $m > n$; that is, we may have any number of beliefs relative to the number of assets that we wish. The only restriction on the set of beliefs that our method requires is that the set of beliefs admits a set of expected returns with a non-empty interior; this rules out the use of certain opposing inequalities to impose equality conditions.

A vector r of expected returns is *consistent* with D if it satisfies the given set of inequality conditions. That is, a consistent return vector is one that could occur given our beliefs. We write

$$\begin{aligned} Q &= \{r \in \mathbb{R}^n | Dr \geq 0\} \\ &= \{r \in \mathbb{R}^n | D_j^T r \geq 0 \text{ for each } j = 1, \ldots, m\} \end{aligned} \tag{4.3}$$

for the set of consistent expected returns. This is a cone in the space $\mathbb{R}^n$ of all possible expected returns, and it is the natural generalization of equation (4.1) to the case of inequality information. Any vector $r \in Q$ may be the actual expected return vector.

A straightforward generalization of the classic construction in Section 4.2.1 would now assert that $v \succeq w$ if and only if $v^{\mathrm{T}} r \geq w^{\mathrm{T}} r$ for all $r \in Q$. However, it will turn out that this is *not* the most useful definition since it brings in the orthogonal components.

We now give a simple example. This is the motivation for our entire work, but the real power of our approach is illustrated by the rich variety of examples in Section 4.4.

Complete sort

The simplest example is that of a *complete sort*, where we have sorted stocks so that

$$r_1 \geq r_2 \geq \cdots \geq r_n$$

We have $m = n-1$ beliefs of the form $r_j - r_{j+1} \geq 0$ for $j = 1, \ldots, n-1$. The belief vectors are of the form $D_j = (0, \ldots, 0, 1, -1, 0, \ldots, 0)^{\mathrm{T}}$, and the matrix D is (empty spaces are zeros)

$$D = \begin{pmatrix} D_1^{\mathrm{T}} \\ \vdots \\ D_m^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} 1 & -1 & & \\ & 1 & -1 & \\ & & \ddots & \ddots \\ & & & 1 \quad -1 \end{pmatrix} \tag{4.4}$$

The consistent cone is a wedge shape in $\mathbb{R}^n$, with a 'spine' along the diagonal $(1, \ldots, 1)$ direction.

4.2.3 The preference relation arising from a portfolio sort

In this section we show that there is a unique preference relation associated with a portfolio sort, naturally extending the preference relation $\succeq$ of Section 4.2.1 to the case of inequality beliefs. Our main tool is an orthogonal decomposition of the return space and of the portfolio space into two linear subspaces. Recalling equation (4.3), we set

$$R^{\perp} = \{r \in \mathbb{R}^n + Dr = 0\} = Q \cap (-Q)$$

$$R = (R^{\perp})^{\perp} = \{w \in \mathbb{R}^n | w^{\perp} r = 0 \text{ for all } r \in R^{\perp}\}$$

By standard linear algebra, $R = \mathrm{Span}(D_1, \ldots, D_m) = \{D^{\mathrm{T}} x \mid x \in \mathbb{R}^m\}$, the subspace spanned by the rows of D. Again, R is the 'relevant' subspace.

For any portfolio w we can again write

$$w = w_0 + w_{\perp} \quad \text{with } w_0 \in R \text{ and } w_{\perp} \in R^{\perp}$$

We want to define preference only in terms of the component w_0 that is relevant to our beliefs, ignoring the orthogonal component $w_{\perp}$. Equivalently, we want to compare portfolios using only components of the return vector for which we have a sign belief, ignoring the perpendicular components which may have either sign.

If w and v are two portfolios, decompose them into parallel parts w_0, $v_0 \in R$ and perpendicular parts $w_{\perp}$, $v_{\perp} \in R^{\perp}$. Then we define

$$v \succeq w \text{ if and only if } v_0^{\mathrm{T}} r \geq w_0^{\mathrm{T}} r \text{ for all } r \in Q.$$

Since any candidate return vector $r = r_0 + r_\perp$ may be similarly decomposed, with $w^{\mathrm{T}} r = w_0^{\mathrm{T}} r_0 + w_\perp^{\mathrm{T}} r_\perp$, an equivalent characterization is

$$v \succeq w \text{ if and only if } v^{\mathrm{T}} r \geq w^{\mathrm{T}} r \text{ for all } r \in Q,$$

where

$$\overline{Q} = Q \cap R$$

That is, it is equivalent whether we test the relevant part of the portfolio weight vector against all consistent returns or the entire weight vector against returns for which we have a sign belief.

We further define *strict* preference as

$$v \succeq w \text{ if and only if } v \succeq w \text{ and } w \not\succeq v,$$

which is equivalent to stating

$$v \succeq w \text{ if and only if } v^{\mathrm{T}} r \geq w^{\mathrm{T}} r \text{ for all } r \in \overline{Q},$$
$$\text{and } v^{\mathrm{T}} r > w^{\mathrm{T}} r \text{ for at least one } r \in \overline{Q}.$$

This notion of preference does *not* mean that portfolio w produces a higher return than portfolio v for every consistent return r, since the portfolios w and v may have different exposure to components of the return vector about which we have no information or opinion, and which may have either sign.

For example, in the complete sort case we write

$$w = \sum_{i=1}^{n-1} x_i D_i + x_n (1, \ldots, 1)^{\mathrm{T}}$$

where $x_1, \ldots, x_n$ are real numbers. The 'relevant' part of our beliefs is spanned by the vectors D_i. A long position in $D_1 = (1, -1, 0, \ldots, 0)^{\mathrm{T}}$ is an investment in the belief that stock 1 has a higher expected return than stock 2; for a universe of n stocks, there are $n-1$ belief vectors. The single remaining dimension is spanned by $(1, \ldots, 1)^{\mathrm{T}}$, a vector which has no significance in the context of our sort information.

This definition is weak since not every pair of portfolios w, v can be compared. If there are some $r \in Q$ for which $w_0^{\mathrm{T}} r > v_0^{\mathrm{T}} r$, and also some $r \in Q$ for which $v_0^{\mathrm{T}} r > w_0^{\mathrm{T}} r$, then neither $v \succeq w$ nor $w \succeq v$. In Section 4.3 we refine the definition to permit such comparisons, but for now we explore the consequences of this rather uncontroversial definition.

4.2.4 *Efficient portfolios*

In MPT investors seek portfolios that provide the maximum level of expected return for a given level of variance. Such portfolios are called efficient. Viewing efficient portfolios as those which are maximally preferable under the preference relation of Section 4.2.1, our goal is now to identify 'maximally preferable' portfolios under the preference relation

arising from a sort, subject to the constraints imposed by risk limits, total investment caps, or liquidity restrictions. These constraints are just as important as the preference definition in constructing optimal portfolios.

In this section we prove two theorems that pave the way for calculating efficient portfolios in the case of portfolio sorts and their generalizations. Although these theorems may seem technically forbidding, they express a familiar idea in finance and optimization: extremal solutions are characterized by the existence of a supporting hyperplane. Because the sets of interest typically do not have smooth surfaces, we need to be somewhat careful with the mathematics.

Let $\mathcal{M} \subset \mathbb{R}^n$ denote the budget constraint set: the set of *allowable* portfolio weight vectors, those that satisfy our constraints. We say that

$$w \text{ is } \textit{efficient} \text{ in } \mathcal{M} \text{ if there is no } v \in \mathcal{M} \text{ with } v \succ w.$$

That is, there is no other allowable portfolio that dominates w, in the sense defined above. This does *not* mean that $w \succeq v$ for all $v \in \mathcal{M}$; there may be many v for which neither $w \succeq v$ nor $v \succeq w$.

We define the efficient set $\hat{\mathcal{M}} \subset \mathcal{M}$ as

$$\hat{\mathcal{M}} = \{w \in \mathcal{M} | w \text{ is efficient in } \mathcal{M}\}$$

The efficient portfolios are far from unique, and in fact for typical constraint sets and belief structures $\hat{\mathcal{M}}$ can be rather large. Nonetheless the construction of the efficient set already gives a lot of information.

We characterize efficient points in terms of the consistent cone Q and the constraint set $\mathcal{M}$. From the example in Section 4.2.5 it is clear that reasonable sets $\mathcal{M}$ are *convex*, but are not typically smooth.

Our main result in this section is the two theorems below: portfolio w is efficient in $\mathcal{M}$ if and only if $\mathcal{M}$ has a supporting hyperplane at w whose normal lies in both the cone Q and the hyperplane R. To give the precise statement we need some definitions. These are based on standard constructions (Boyd and Vandenberghe, 2004), but we need some modifications to properly account for our orthogonal subspaces.

For any set $A \subset \mathbb{R}^n$, we define the *dual* (or *polar*) set

$$A^* = \{x \in \mathbb{R}^n | x^{\mathrm{T}} y \geq 0 \text{ for all } y \in A\}$$

Thus $v \succeq w$ if $v - w \in \overline{Q}^*$, or equivalently, if $v_0 - w_0 \in \overline{Q}^*$. It is also useful to define the interior of the consistent cone

$$Q^\circ = \{r \in \mathbb{R}^n | Dr > 0\}$$

and its planar restriction, the relative interior of $\overline{Q}$ in R,

$$\overline{Q^\circ} = Q^\circ \cap R$$

All of our sets have linear structure along $R^\perp$: we may write $Q = \overline{Q} \oplus R^\perp$ and $Q^\circ = \overline{Q}^\circ \oplus R^\perp$, where $\oplus$ denotes orthogonal sum. As a consequence, $Q^\circ = \varnothing$ if and only if $\overline{Q}^\circ = \varnothing$.

A *normal to a supporting hyperplane for $\mathcal{M}$ at w* is a non-zero vector $b \in \mathbb{R}^n$ such that $b^{\mathrm{T}}(v-w) \leq 0$ for all $v \in \mathcal{M}$. A *strict normal* is one for which $b^{\mathrm{T}}(v-w) < 0$ for all $v \in \mathcal{M}$ with $v \neq w$.

Now we can state and prove our two theorems that characterize the relationship between normals and efficiency:

1. Theorem 1. Suppose that $\mathcal{M}$ has a supporting hyperplane at w whose normal $b \in \overline{Q}$. If b is a strict normal, or if $b \in \overline{Q}^{\circ}$, then w is efficient.
2. Theorem 2. Suppose that Q has a non-empty interior, and suppose that $\mathcal{M}$ is convex. If w is efficient in $\mathcal{M}$, then there is a supporting hyperplane to $\mathcal{M}$ at w whose normal $b \in \overline{Q}$.

Since we have not assumed smoothness, these theorems apply to non-smooth sets having faces and edges. We defer the proofs of these theorems to Appendix A. In the next section, we show how to determine the efficient sets explicitly for the examples of most interest.

Connection to classic theory

Suppose we are given an expected return vector ρ. In our new framework, we introduce the single belief vector $D_1 = \rho$. We are thus saying that we believe the actual return vector may be any $r \in Q$; that is, we only believe that $\rho^{\mathrm{T}} r \geq 0$. This is a weaker statement than the belief that $r = \rho$. However, the procedure outlined here tells us to ignore the orthogonal directions, about which we have no sign information. At any efficient point we have a normal $b \in \overline{Q}$; that is, $b = \lambda\rho$ for some $\lambda > 0$. This is exactly the same result that we would have obtained in the classic formulation, and our new formulation extends it to more general inequality belief structures.

Complete sort

For constructing examples, it is useful to have an explicit characterization of $\overline{Q}$ as a positive span of a set of basis vectors. That is, we look for an $n \times m$ matrix E so that $\overline{Q} = \{Ex | x \geq 0 \text{ in } \mathbb{R}^m\}$.

In general, finding the columns $E_1, \ldots, E_m$ of E is equivalent to finding a convex hull. But if $D_1, \ldots, D_m$ are linearly independent, which of course requires $m \leq n$, then E may be found as the Moore–Penrose pseudo-inverse of D: Span $(E_1, \ldots, E_m) = \text{Span}(D_1, \ldots, D_m)$ and $E_i^{\mathrm{T}} D_j = \delta_{ij}$.

For a single sort, the dual of the $(n-1) \times n$ matrix D from (4) is the $n \times (n-1)$ matrix

$$E = \frac{1}{n} \begin{pmatrix} n-1 & n-2 & n-3 & \ldots & 2 & 1 \\ -1 & n-2 & n-3 & \ldots & 2 & 1 \\ -1 & -2 & n-3 & \ldots & 2 & 1 \\ \vdots & \vdots & \ddots & & \ddots & \vdots \\ -1 & -2 & -3 & \ldots & -(n-2) & 1 \\ -1 & -2 & -3 & \ldots & -(n-2) & -(n-1) \end{pmatrix} \tag{4.5}$$

For $n = 3$, the difference vectors and their duals are

$$D_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix},\ D_2 = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix},\ E_1 = \frac{1}{3}\begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix},\ E_2 = \frac{1}{3}\begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}$$

so that $D_i^{\mathrm{T}} E_j = \delta_{ij}$. The angle between D_1 and D_2 is 120°, the angle between E_1 and E_2 is 60°, and they all lie in the plane R whose normal is $(1, 1, 1)^{\mathrm{T}}$, the plane of the image in Figure 4.1.

4.2.5 Examples of constraint sets

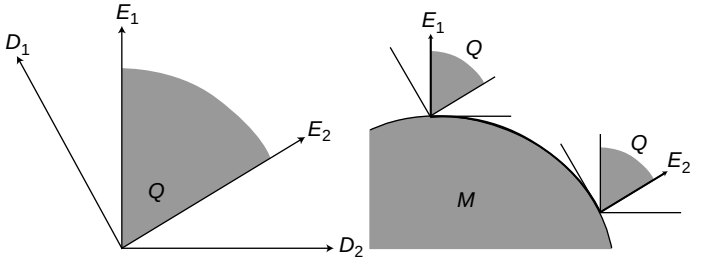
We now demonstrate the computation of efficient sets in the most common situations that arise in the practice of investment management: when portfolio constraint sets are based on a total risk budget or on a maximum investment constraint, with possible additional constraints from market neutrality or transaction costs. As we have noted, the efficient sets are defined by two independent but equal pieces of input:

1. The belief vectors D about the expected return vector r.
2. The constraint set $\mathcal{M}$ imposed on the portfolio w.

We will consider several different structures for $\mathcal{M}$, and within each we will construct the efficient set for the complete sort example. In Section 4.4 we consider more general beliefs.

Total investment constraint

In our first example, we are limited to total investment at most W dollars, with long or short positions permitted in any asset. We take $\mathcal{M} = \{w \in \mathbb{R}^n \,||\, w_1| + \cdots + |w_n| \le W\}$.



The view is from the direction of $(1, 1, 1)^{\mathrm{T}}$, so the image plane is R. The left panel is in the space of expected return r, where Q is the consistent set; the full three-dimensional shape is a wedge extending perpendicularly to the image plane. The right panel shows a smooth constraint set M of portfolio vectors w; the efficient set is the shaded arc, where the normal is in the positive cone of E_1, E_2. Along this arc, the normal must be in the image plane; if M is curved in three dimensions, then the efficient set is only this one-dimensional arc.

Figure 4.1 Geometry for three assets, with two sorting conditions.

We first consider the case of a single sorted list. Start with a portfolio weighting $w = (w_1, \ldots, w_n)$. If $w_j > 0$ for some $j = 2, \ldots, n$, then form the portfolio $w' = (w_1 + w_j, \ldots, 0, \ldots, w_n)$, in which the component w_j has been set to zero by moving its weight to the first element. This has the same total investment as w if $w_1 \geq 0$, and strictly less if $w_1 < 0$. It is more optimal since the difference $w' - w = (w_j, \ldots, -w_j, \ldots) = w_j(D_1 + \ldots + D_{j-1})$ is a positive combination of difference vectors.

Similarly, if any $w_j < 0$ for $j = 1, \ldots, n-1$, we define a more optimal portfolio $w' = (w_1, \ldots, 0, \ldots, w_n + w_j)$, which has the same or less total investment, and is more optimal than w.

We conclude that the only possible efficient portfolios are of the form $w = (w_1, 0, \ldots, 0, w_n)$ with $w_1 \geq 0$, $w_n \leq 0$, and $|w_1| + |w_n| = W$, and it is not hard to see that all such portfolios are efficient. This is the classic portfolio of going maximally long the most positive asset, and maximally short the most negative asset. In this example, the covariance matrix plays no role.

By similar reasoning, the efficient portfolios in the case of multiple sectors go long the top asset and short the bottom asset within each sector; any combination of overall sector weightings is acceptable.

Risk constraint

Here we take $\mathcal{M} = \{w \in \mathbb{R}^n | w^{\mathrm{T}} V w \leq \sigma^2\}$, where V is the variance–covariance matrix of the n assets and σ is the maximum permissible volatility. This set is a smooth ellipsoid, and at each surface point w it has the unique strict normal $b = Vw$ (up to multiplication by a positive scalar). Conversely, given any vector $b \in \mathbb{R}^n$, there is a unique surface point w in $\mathcal{M}$ having normal b; w is a positive multiple of $V^{-1}b$. As noted above, b is in effect a vector of imputed returns, and any such vector corresponds to exactly one efficient point on $\mathcal{M}$.

By the theorems, $w \in \mathcal{M}$ is efficient if and only if $b \in \overline{Q}$. So we may parameterize the set $\hat{\mathcal{M}}$ of efficient points by $b = Ex$ with $x \in \mathbb{R}^m$ with $x \geq 0$. We may write this explicitly as

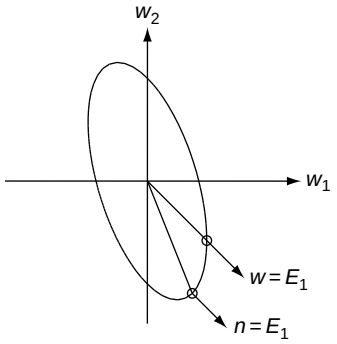
$$w = V^{-1} \sum_{j=1}^{m} x_j E_j$$

with appropriate scaling so $w^{\mathrm{T}} V w = \sigma^2$.

The efficient set $\hat{\mathcal{M}}$ is a portion of the surface of the risk ellipsoid intersected with the plane where the local normal is in $\overline{Q}$. For example, in the case of a single sort it is a distorted simplex with $n-1$ vertices corresponding to only a single x_j being non-zero. In general, each of the $n!$ possible orderings gives a different set $\hat{\mathcal{M}}$, and the set of these possibilities covers the whole set $\mathcal{M} \cap V^{-1}\overline{Q}$. That is, the size of $\hat{\mathcal{M}}$ is $1/n!$ of the entire possible set.

To select a single optimal portfolio we must pick one point within this set. For example, we might take the 'centre' point with $x = (1, \ldots, 1)$, which gives $b = E_1 + \ldots + E_m$. In the case of a single sorted list, this gives the linear return vector

$$b_i = \sum_{j=1}^{n-1} E_{ij} = \frac{n+1}{2} - i$$



The lower solution in this diagram is the unique efficient point with no constraint of market neutrality; the upper solution is the market-neutral solution. The risk ellipsoid takes $\sigma_1 = 2\sigma_2$ and $\rho = 0.5$.

Figure 4.2 Optimal portfolios for two assets.

Of course, this is to be multiplied by V^{-1} and scaled to get the actual weights. In Section 4.3 we propose a more logical definition of 'centre' point, and in Section 4.5 we demonstrate that the difference is important.

Figure 4.2 shows the unique efficient portfolio in the case of two assets; since there is only one vector E_1 there is only a single point.

Risk constraint with market neutrality

Suppose that we impose two constraints. We require that the portfolio has a maximum level of total volatility as described above. In addition, we require that the portfolio be *market-neutral*, meaning that $\mu^{\mathrm{T}} w = 0$, where μ is a vector defining the market weightings. We assume that $\mu \notin \mathrm{Span}(D_1, \ldots, D_m)$. For example, for equal weightings, $\mu = (1, \ldots, 1)^{\mathrm{T}}$.

The set $\mathcal{M}$ is now an ellipsoid of dimension $n-1$. At 'interior' points where $w^{\mathrm{T}} V w < \sigma^2$, it has normal $\pm\mu$. At 'boundary' points where $w^{\mathrm{T}} V w = \sigma^2$, it has a one-parameter family of normals $\mathcal{B} = \{\alpha V w + \beta\mu \mid \alpha \geq 0,\ \beta \in \mathbb{R}\}$.

Proof: We need to show that $b^{\mathrm{T}}(w - v) \geq 0$ for all $b \in \mathcal{B}$ and all $v \in \mathcal{M}$. But

$$\begin{aligned} b^{\mathrm{T}}(w - v) &= \alpha(w^T V w - w^T V v) + \beta(\mu^T w - \mu^T v) \\ &= \frac{1}{2}\alpha(w - v)^T V(w - v) + \frac{1}{2}\alpha(w^T V w - v^T V v) \geq 0 \end{aligned}$$

since $\mu^{\mathrm{T}} w = \mu^{\mathrm{T}} v = 0$, $w^{\mathrm{T}} V w = \sigma^2$, $v^{\mathrm{T}} V v \leq \sigma^2$, and V is positive definite. It is clear that these are the only normals since the boundary of $\mathcal{M}$ has dimension $n-2$. The strict normals to $\mathcal{M}$ at w are those with $\alpha > 0$.

For w to be efficient, we must have $\mathcal{B} \cap \overline{Q} \neq \varnothing$. That is, there must exist $\alpha \geq 0$ and β not both zero and $x_1, \ldots, x_m \geq 0$, not all zero, so that

$$\alpha V w + \beta \mu = x_1 E_1 + \cdots + x_m E_m$$

Since $\mu \notin \text{Span}(E_1, \ldots, E_m)$, we must have $\alpha > 0$, and this is equivalent to

$$V w + x_1 E_1 + \cdots + x_m E_m + y\mu$$

where y is determined so that $\mu^{\mathrm{T}} w = 0$. We may explicitly parameterize the set of efficient w by

$$w = x_1 V^{-1} E_1 + \cdots + x_m V^{-1} E_m + y V^{-1} \mu$$

with

$$y = -\frac{x_1 \mu^{\mathrm{T}} V^{-1} E_1 + \cdots + x_m \mu^{\mathrm{T}} V^{-1} E_m}{\mu^{\mathrm{T}} V^{-1} \mu}$$

or

$$w = V^{-1} \sum_{j=1}^{m} x_j \tilde{E}_j, \quad \tilde{E}_j = E_j - \frac{\mu^{\mathrm{T}} V^{-1} E_j}{\mu^{\mathrm{T}} V^{-1} \mu} \mu$$

As $x_1, \ldots, x_m$ range through all non-negative values, this sweeps out all efficient w, with suitable scaling to maintain $w^{\mathrm{T}} V w = \sigma^2$. As with the previous case, this is a rather large efficient family, but in the next section we will show how to choose a single optimal element. Figure 4.2 shows the extremely simple case of two assets.

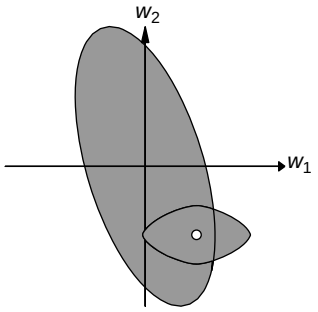
Transaction cost limits

An extremely important issue in practice is the transaction costs that will be incurred in moving from the current portfolio to another one that has been computed to be more optimal. If portfolios are regularly rebalanced, then the holding period for the new portfolio will be finite, and the costs of the transition must be directly balanced against the expected increase in rate of return.

One common way to formulate this trade-off follows the formulation of volatility above: a rigid limit is given on the transaction costs that may be incurred in any proposed rebalancing, and the efficient portfolios are sought within the set of new portfolios that can be reached from the starting portfolio without incurring unacceptable costs. In this formulation, our procedure naturally incorporates transaction cost modelling.

Let w^0 be the current portfolio, and w be a candidate new portfolio. In order to rebalance w^0 to w, $w_i - w_i^0$ shares must be bought in the ith asset, for $i = 1, \ldots, n$; if this quantity is negative, then that number of shares must be sold.

A popular and realistic model (Almgren and Chriss, 2000; Almgren, 2003) for market impact costs asserts that the cost per share of a trade execution schedule is proportional to the kth power of the 'rate of trading' measured in shares per unit time, where k is some



The ellipsoid represents the total risk limit, with $\sigma_1/\sigma_2 = 2$ and $\rho = 0.5$. The curved diamond represents the new portfolios that are reachable from the given starting portfolio with a limit on total transition cost; $\eta_2/\eta_1 = 2.5$, and the exponent $k = 1/2$. The dark shaded region is the set of new portfolios that satisfy both constraints.

Figure 4.3 Transaction cost limit in combination with risk limit, for two assets.

positive exponent; $k = 1/2$ is a typical value. Assuming that the programme is always executed in a given fixed time, and recalling that the pershare cost is experienced on $w_i - w_i^0$ shares, the total cost of the rebalancing is

$$\text{Rebalancing cost} \equiv F(w) = \sum_{i=1}^{n} \eta_i |w_i - w_i^0|^{k+1}$$

where η_i is a stock-specific liquidity coefficient (we neglect 'cross-impacts', where trading in stock i affects the price received on simultaneous trades in stock j).

If a total cost limit C is imposed, then the constraint set becomes

$$\mathcal{M} = \mathcal{M}_0 \bigcap \{w \in \mathbb{R}^n | F(w) \leq C\}$$

where $\mathcal{M}_0$ is a pre-existing constraint set that may take any of the forms described above, such as total risk limit. Since $k > 0$, F is a convex function and hence its level sets are convex. Since M_0 is assumed convex, the intersection M is also a convex set, and the theorems above apply. Computing the efficient set is then a non-trivial problem in mathematical programming, though for the important special case $k = 1/2$, methods of cone programming may be applied. The geometry is illustrated in Figure 4.3. Note that the intersection does not have a smooth boundary, although each individual set does; in the case shown here, efficient portfolios will most likely be at the bottom right corner of the dark region.

4.3 Optimal portfolios

In this section we continue to pursue the idea that a rational investor will prefer to hold a portfolio that delivers the highest expected return for a given level of risk, but now we extend the notion of preference to include the possibility that one portfolio will deliver a higher expected return *more often* than another. Thus we extend the preference relation of

the previous section to produce a unique efficient portfolio that is economically superior to the others in a certain well-defined sense.

The previous section's preference relation ranks one portfolio relatively higher than another if its expected return is greater than the other's across all expected returns consistent with the portfolio sort. The previous section completely analyzes the situation in which an investor prefers one portfolio to another when it is 100% certain that the portfolio will have a higher expected return. This section develops the theory further by providing a methodology for choosing among the portfolios within the efficient set.

To accomplish this, we refine the preference relation on portfolios by relaxing the requirement of 100% certainty and positing that a rational investor will prefer one portfolio to another if the portfolio delivers a greater expected return a greater percentage of the time. Of course, the challenge is to put meaning to the notion of one portfolio delivering a higher expected return than another a greater percentage of the time. To do this we are forced to study probability measures on the consistent cone to capture the inherent uncertainty associated with what the expected return is.

To develop the theory in this section we need to study in detail the geometry of the consistent cone when it is equipped with a probability measure. While the methodology requires some unfortunate technicalities, we stress that the results are simple and intuitive in nature and in practice easy to compute. This leads us to believe that there are simpler and easier ways to arrive at the same results that we have not yet been able to find. That said, given the technical nature of the proofs, we have relegated them to Appendix A.

4.3.1 Outline of approach

The basic outline of our approach is as follows. We start by equipping the consistent cone with any probability distribution from a broad class of probability distributions that possess certain symmetry conditions. This allows us to quantify the subjective likelihood of any of a range of expected returns being the actual expected return. We show that by using a specific distribution we can produce a preference relation among efficient portfolios merely by stating that a rational investor would always choose between two portfolios by selecting the one that delivers a higher expected return a greater percentage of the time.

This preference relation is mathematically coherent, it is economically rational, and it computes the same optimal portfolio for all the distributions from the class we set forth. Further, we show that the optimal portfolio is straightforward to compute in practice because the preference relation defined is completely captured by a linear function on the space of portfolios. By this we simply mean that there is a linear function c on the set of portfolios written $w \cdot c^{\mathrm{T}}$ such that we prefer portfolio w to portfolio v if and only if $(w–v) \cdot c^{\mathrm{T}} > 0$. In this way, finding the optimal portfolio comes down to two problems: first, computing c; second, finding the maximum among a constrained set of portfolios of $w \cdot c^{\mathrm{T}}$.

The first problem turns out to be straightforward because we show that the linear function c is exactly the same as the geometric centroid (that is, the centre of gravity with respect to the measure) of the consistent cone. We study the centroid and its properties in detail in the next section. Moreover, although the centroid is a completely

geometric notion, in practice it is straightforward to compute and we provide algorithms for computing it in several different contexts.

In this guise, computing optimal portfolios reduces to maximizing a known linear function over the space of portfolios meeting whatever set of constraints we have. Thus all of the technical theory falls away and we are back in a situation quite common to practitioners. We have a set of assets, a known order of the expected returns of the assets, and an algorithm for producing a linear function whose maximum over the set of feasible portfolios is the economically optimal portfolio for this situation.

4.3.2 The centroid

The fundamental idea behind our refining of the previous preference relation is to abandon the idea of comparing only those portfolios for which all expected returns are better and replace it with the weaker idea that *more* expected returns are better. The latter notion clearly applies to all portfolios once we make sense of what 'more' means. To do this requires placing a suitably defined probability measure on the set of consistent expected returns. We can then define that one portfolio is preferred to another precisely when the measure of the set where the one has greater expected returns than the other is greater than where the other has greater expected returns. We then prove the remarkable fact that for a rather broad class of measures this preference relation is defined precisely by the linear function defined by the centre of mass (i.e. the centroid) of the consistent set.

Extended preference definition

Consider two portfolios w and v, along with their 'relevant' parts w_0 and v_0 as defined in Section 4.2.3. In general, unless $w \succeq v$ or $v \succeq w$, there will be some consistent expected returns for which w_0 has an expected return superior to v_0, and a complementary set of expected return vectors for which v_0 gives a higher expected return. Only if $w \succeq v$ or $v \succeq w$ do all consistent expected return vectors give better results for one portfolio or the other.

To define the extended preference relation we need to introduce some notation. Although the portfolio vectors and the return vectors are both elements of the vector space $\mathbb{R}^n$, it will be convenient to denote the space of return vectors by $\mathcal{R}$ and the space of portfolio vectors by $\mathcal{W}$; the inner product $w^{\mathrm{T}}r = \langle w, r\rangle$ defines a bilinear map $\mathcal{W} \times \mathcal{R} \to \mathbb{R}$. For any portfolio vector $w \in \mathcal{W}$, we define $Q_w \subset Q$ by

$$Q_w = \{r \in Q | w_0^{\mathrm{T}} r \geq 0\}$$

Clearly, Q_w and Q_{-w} are complementary in the sense that $Q_w \cup Q_{-w} = Q$.

We could equally well formulate the comparison by using a cone $\overline{Q}_w = \{r \in \overline{Q} | w^{\mathrm{T}} r \geq 0\}$. These formulations will be equivalent under the mirror symmetry assumption below, but for now we continue as we have started.

For two portfolios w and v, we now consider the complementary sets $Q_{w_0 - v_0}$ and $Q_{v_0 - w_0}$. Here $Q_{w_0 - v_0}$ is the set of consistent return vectors for which the portfolio w will be at least as large an expected return as v, comparing only the relevant parts w_0, v_0.

We now define the extended preference relation, retaining the same notation $v \succeq w$ as meaning v is preferred to w.

> Let μ be a probability measure on $\mathcal{R}$ so that $\mu(Q) = 1$. We say that $v \succeq w$ (with respect to μ) if $\mu(Q_{v_0 - w_0}) \geq \mu(Q_{w_0 - v_0})$.

This definition includes and extends the definition of the preference relation of the previous section. Stated in these terms, that preference relation said that $v \succeq w$ if $\mu(Q_{v_0 - w_0}) = 1$.

> We write $v \simeq w$ if $\mu(Q_{v_0 - w_0}) = \mu(Q_{w_0 - v_0})$ and we write $v \succ w$ if $\mu(Q_{v_0 - w_0}) > \mu(Q_{w_0 - v_0})$.

The above definition is fairly broad, as it defines one preference relation for every probability measure on Q. Below we make a specific choice for the measure μ. But for now we assume that a density has been given, and we follow through on some of the geometrical consequences.

The centroid

In order to find efficient portfolios with the new preference relation, we will identify a real-valued function $h(w)$ on the space of portfolios $\mathcal{W}$ such that $w \succ v$ if and only if $h(w) > h(v)$. Then the maximizer of this function on a given convex budget set is the unique efficient portfolio under our preference relation.

To identify the function $h(w)$, we consider its level sets, defined by the relation $w \simeq v$ for a given portfolio v. Any such w must satisfy the condition

$$\mu(Q_{w_0 - v_0}) = \mu(Q_{v_0 - w_0}) = \frac{1}{2}$$

where μ is the measure defined above.

To understand this properly we look at the space $\mathcal{H}$ of hyperplanes through the origin in $\mathcal{R}$. For $w \in \mathcal{W}$ define $w^{\perp} \subset \mathcal{R}$ by

$$w^{\perp} = \{r \in \mathcal{R} | w^{\mathrm{T}} r = 0\}$$

In this way each $w \in \mathcal{W}$ defines a hyperplane H through the origin in $\mathcal{R}$ whose normal is w. Let $\mathcal{H}$ be the set of all such hyperplanes. Conversely, for any $H \in \mathcal{H}$, there is a one-parameter family of normals w so that $H = w^{\perp}$: if w is a normal to H then so is λw for any $\lambda \in \mathbb{R}$.

Now, for a given hyperplane $H = w^{\perp} \in \mathcal{H}$, we say that H *bisects* the set of consistent returns Q if and only if

$$\mu(Q_w) = \mu(Q_{-w}) = \frac{1}{2}$$

Clearly, $w \simeq v$ if and only if the hyperplane $(w_0 - v_0)^{\perp}$ bisects Q.

Let $\mathcal{P}$ be the subset of $\mathcal{H}$ of all hyperplanes through the origin that bisect Q. Let c be the *centroid* or *centre of mass* of Q; that is, the mean of all points in Q under the measure μ as defined by the integral

$$c = \int_{r \in Q} r \mathrm{d}\mu$$

The following standard result characterizes $\mathcal{P}$ in terms of c.

Theorem 3: The line joining the origin and the centroid is the intersection of all hyperplanes through the origin in $\mathcal{R}$ that bisect Q:

$$\{\lambda c | \lambda \in \mathbb{R}\} = \bigcap_{H \in p} H$$

Since we are only interested in rays rather than points, we often say that c 'is' this intersection.

We now make the following assumption about the measure μ:

μ has mirror symmetry about the plane $\overline{Q}$.

As a consequence, $c \in \overline{Q}$ and we have the minor

Lemma: For any $w \in \mathcal{W}$, $w_c^{\mathrm{T}} = w_0^{\mathrm{T}} c$.

Proof: Write $w = w_0 + w_\perp$ and observe that $w_\perp^{\mathrm{T}} c = 0$ since $c \in \overline{Q}$.

The symmetry assumption is natural given our lack of information about return components orthogonal to our belief vectors, and it leads immediately to a characterization of our portfolio preference relation in terms of the centroid.

Theorem 4: Let w, $v \in \mathcal{W}$ be portfolios and c be the centroid vector as defined above. Then we have

$$w \simeq v \Leftrightarrow (w - v)c^{\mathrm{T}} = 0$$

Proof: By definition, $w \simeq v$ if and only if $\mu(Q_{w_0 - v_0}) = \mu(Q_{v_0 - w_0})$. This means precisely that $(w_0 - v_0)^\perp$ must bisect Q. That is, $(w_0 - v_0)^\perp \in \mathcal{P}$. This implies that $c \in (w_0 - v_0)^\perp$, or in other words, $(w_0 - v_0)^{\mathrm{T}} c = 0$. By the lemma, this is equivalent to $(w - v)^{\mathrm{T}} c = 0$.

Conversely, if $(w - v)^{\mathrm{T}} c = 0$ then $(w_0 - v_0)^{\mathrm{T}} c = 0$; this means that $c \in (w_0 - v_0)^\perp$, which implies by Theorem 1 that $(w_0 - v_0)^\perp \in \mathcal{P}$. That is, that $w \simeq v$.

4.3.3 Centroid optimal portfolios

The theorems of the previous section characterize our portfolio preference relation in terms of the centroid vector c. This means that the entire problem of calculating efficient or optimal portfolios is reduced to manipulations involving the centroid vector: two portfolios are equivalently preferable if and only if $w^{\mathrm{T}}c = v^{\mathrm{T}}c$. We now cast the notion of efficiency in terms of the centroid vector c.

Given a convex budget set $\mathcal{M}$, a point $w \in \mathcal{M}$ is *efficient* (in the sense that there is no portfolio in $\mathcal{M}$ preferred to it) if and only if $(v - w)^{\mathrm{T}}c \leq 0$ for all $v \in \mathcal{M}$. That is, $\mathcal{M}$ must have a supporting hyperplane at w whose normal is c; as described in Section 4.2.4. We now summarize this in a formal definition.

Definition 1: Let c be the centroid vector related to a portfolio sort. Let $\mathcal{M} \subset \mathcal{W}$ be a convex budget constraint set. A candidate portfolio $w \in \mathcal{M}$ is *centroid optimal* if there is no portfolio $v \in \mathcal{M}$ such that $v^{\mathrm{T}}c > w^{\mathrm{T}}c$.

We now state our main result which allows us to calculate centroid optimal portfolios in practice.

Theorem 5: If w is centroid optimal, then $\mathcal{M}$ has a supporting hyperplane at w whose normal is c. Hence, by Theorem 4, it is efficient with respect to the equivalence relation defined in Section 4.2.3.

In order for w to be efficient in the sense of Section 4.2.4, we must further require that $c \in \overline{Q} = Q \cap R$. This will be true if and only if the density μ is *symmetric* about the plane R; symmetry in this sense is part of our assumptions in the next section.

Just as in classic portfolio theory (Section 4.2.1), the magnitude of the vector c has no effect on the resulting optimal portfolio, given a specified budget constraint set. In effect, we consider the centroid c to be defined only up to a scalar factor; we think of it as a *ray* through the origin rather than a single point. In Appendix B we present efficient techniques for computing the centroid vector c, both for single portfolios and for collections of sectors.

The most common budget constraint is a risk constraint based on total portfolio variance as in Section 4.2.5, 'Risk constraint'. If the portfolio constraints are of more complicated form involving, for example, position limits, short sales constraints, or liquidity costs relative to an initial portfolio, then all the standard machinery of constrained optimization may be brought to bear in our situation. Constraints on the portfolio weights are 'orthogonal' to the inequality structure on the expected returns.

4.3.4 Symmetric distributions

The above work refines the portfolio preference relation of Section 4.2 to yield a unique optimal portfolio in terms of the centroid vector of the consistent cone. Our refined preference relation and the centroid depend on the specification of the distribution of expected returns, which introduces an element of parametrization into the formulation of our problem in that it characterizes how the expected returns consistent with a sort are likely to be distributed.

What is the correct probability distribution on the space of consistent expected returns? By hypothesis, we have no information about the expected return vector other than the inequality constraints that define the consistent set (recall that we believe that the covariance structure is not related to the expected moments). This forces us to make the most 'neutral' possible choice.

We assume that the probability density μ is *radially symmetric* about the origin, restricted to the consistent cone Q. A radially symmetric density is one which is the same along any ray from the origin. That is, we consider densities that can be written in the form $\mu(r) = f(|r|)g(r/|r|)$ where $f(\rho)$ for $\rho \geq 0$ contains the radial structure and $g(\omega)$ with $|\omega| = 1$ contains the azimuthal structure. We then require that $g(\omega)$ be a constant density, restricted to the segment of the unit sphere included in the wedge Q. The radial function $f(\rho)$ may have any form, as long as it decreases sufficiently rapidly as $\rho \to \infty$ so that the total measure is finite.

For example, the n-component uncorrelated normal distribution, with density proportional to $f(\rho) = \rho^{n-1} \exp(-\rho^2/2R^2)$, is a candidate distribution. Or, we could choose a distribution uniform on the sphere of radius R. In both of these examples, R is a typical scale of return magnitude, for example 5% per year, and may have any value.

An essential feature of our construction is that we do not need to specify the value of the radius R or even the structure of the distribution: the relative classification of returns is identical under *any* radially symmetric density. This will be apparent from the construction below, and mirrors the observation in Section 4.2.1 that with a fixed risk budget, the classic mean-variance portfolio depends only on the direction of the expected return, not on its magnitude. In effect, since all the sets of interest to us are cones, we measure their size by their angular measure.

Radial symmetry means geometrically that two points in return space $\mathcal{R}$ have equal probability if they have the same Euclidean distance from the origin: $|r|^2 = r_1^2 + \cdots + r_n^2$. It is not obvious that this distance is the most appropriate measure; one might argue that the metric should depend somehow on the covariance matrix. But, as we have argued in Section 4.2, we assume that our information about the first moments of the return distribution is independent of the second moments. Thus we propose this as the only definition that respects our lack of information aside from the homogeneous inequality constraints.

4.3.5 Computing the centroid

The key fact about the centroid vector is that we capture a rather complicated equivalence relation for portfolios in a very simple geometric construction. We have shown that if we have a formula for the centroid then we transform the problem of finding efficient portfolios into a linear optimization problem, which is solvable by known means. Section 4.3.2 gives us the key to calculating the centroid for an arbitrary portfolio sort via a straightforward Monte Carlo approach.

In Appendix B we demonstrate how to do this Monte Carlo simulation in some detail. The key observation about this approach is that it is straightforward and directly related to Section 4.3.2. The entire trick to the computation is providing a method for randomly sampling from the consistent cone Q associated to a given sort. This method, therefore, works in principle for any sort, whether it be a complete sort, partial sort, or even the cone associated to multiple sorts (see Section 4.4 for more on this).

In the case of a complete sort, the method for sampling from the consistent cone Q boils down to generating a draw from an uncorrelated n-dimensional Gaussian and then sorting the draw according to the sort. This sorting process is equivalent to applying a sequence of reflections in $\mathbb{R}^n$ that move the draw into the consistent cone. The Monte Carlo simulation is then the process of averaging of these draws which in effect computes the integral in Section 4.3.2.

In the case of a complete sort the averaging process is equivalent to drawing from the order statistics of a Gaussian. The component associated to the top-ranked stock is a draw from the average of the largest draw from n independent draws from a Gaussian, the next ranked stock is the average of the second largest draw from n independent draws from a Gaussian, etc. We show in Appendix B that this procedure produces something very close to the inverse image of a set of equally spaced points of the cumulative normal function (see Section 4.4 for a picture of what this looks like).

Thus, centroid optimal portfolios in the case of a complete sort are equivalent to portfolios constructed by creating a set of expected returns from the inverse image of the cumulative normal function, where the top-ranked stock receives the highest alpha, and the alphas have the same order as the stocks themselves. So, the centroid optimal portfolio is the same portfolio as the Markowitz optimal portfolio corresponding to a set of expected returns that are *normally* distributed in the order of the corresponding stocks. This is remarkable in light of the completely general framework from which this fact was derived. In a completely natural, economic way, the optimal portfolio to trade is exactly that portfolio derived in the Markowitz framework from a set of normally distributed expected returns.

However, because our approach in this chapter is completely general and applies to any portfolio sort, we can apply the method to a much broader set of sorts than simply complete sorts. In the next section we examine these in detail.

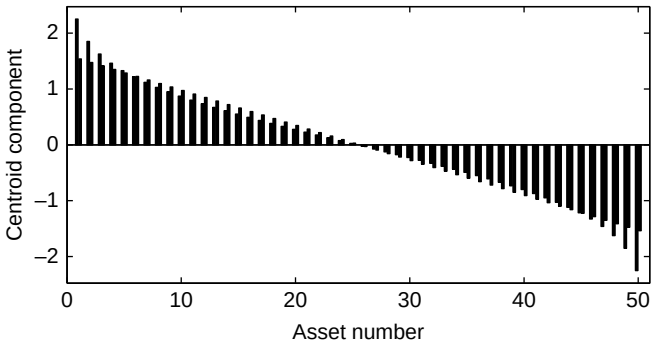
4.4 A variety of sorts

Above we outlined in detail how to calculate centroid optimal portfolios. Our construction was completely general. We showed that, given the equivalence relation which states that we prefer one portfolio to another when the one's expected return is greater than the other's (on its set of relevant direction) for returns consistent with the ordering *more often* (with respect to a certain measure μ) than it is not greater. We showed that this equivalence relation is completely characterized by the linear function given by the centroid of the cone of consistent returns Q with respect to μ.

The purpose of this section is to illustrate the variety of inequality criteria to which our methodology can be applied, and to show the centroid in all these cases.

4.4.1 Complete sort

To begin, we illustrate the simple sort used as an example in Section 4.2. Figure 4.4 shows the centroid vector, for a moderate portfolio of 50 assets, compared to the linear weighting. These vectors are defined only up to a scalar constant; for the plot they have been scaled to have the same sum of squares. As suggested in Section 4.2.5, 'Risk restraint',



The centroid overweights very high- and very low-ranked stocks while underweighting the middle. The weights have been chosen so that the sum of the absolute value of each are the same.

Figure 4.4 The centroid vector for a complete sort of 50 assets, compared with linear weights.

the linear portfolio is a natural way to smoothly weight the portfolio from the first (believed best) asset to the the last (believed worst).

The linear portfolio in effect assigns equal weight to each difference component. By comparison, the centroid portfolio curves up at the ends, assigning greater weight to the differences at the ends than the differences in the middle of the portfolio. The reason for this is that typical distributions have long 'tails', so two neighbouring samples near the endpoints are likely to be more different from each other than two samples near the middle. In fact, the centroid profile looks like the graph of the inverse cumulative normal distribution; this is indeed true when n is reasonably large, and is exploited in Appendix B.

4.4.2 Sector sorts

Next, we look at the case where we assume that each stock in our universe is assigned to a distinct sector and that within each sector we have a complete sort. If we have k sectors with m_i stocks in sector i then we can order our stocks as follows:

$$(r_1, r_2, \cdots, r_{n_1}), (r_{n_1}+1, \cdots, r_{n_2}), \cdots, (r_{n_{k-1}}+1, \cdots, r_n)$$

with $n_1 = m_1, n_2 = m_1 + m_2, \ldots, n_k = m_1 + \cdots + m_k = n$. Then we assume a sort within each group:

$$r_1 \geq \cdots \geq r_{n_1},\ r_{n_1+1} \geq \cdots \geq r_{n_2}, \cdots,\ r_{n_{k-1}+1} \geq \cdots \geq r_n$$

This is almost as much information as in the complete sort case except that we do not have information about the relationships at the sector transitions. If there are k sectors, there are $m = n-k$ columns D_j of the form $(0, \ldots, 0,\ 1,\ -1,\ 0, \ldots, 0)^{\mathrm{T}}$, and the matrix D is of size $(n-k) \times n$. The consistent cone Q is a Cartesian product of the sector cones of dimension $m_1, \ldots, m_k$.

As a specific example, if there are five assets, divided into two sectors of length two and three, then

$$D = \left(\begin{array}{cc|ccc} 1 & -1 & & & \\ & & & & \\ \hline & & 1 & -1 & \\ & & & 1 & -1 \end{array}\right)$$

Orthogonal decomposition For k sectors, there are k orthogonal directions, corresponding to the mean expected returns within each sector. $R^{\perp}$ has dimension k, and R has dimension $n-k$.

Matrix structure The dual matrix is multiple copies of the single-sector dual in equation (4.5).

Centroid profile Figure 4.5 shows the centroid portfolio for two sectors: one sector has ten assets and the other has fifty assets, for a total portfolio size of $n = 60$. Within each sector, the vector is a scaled version of the centroid vector for a single sector. Although the overall scaling of the graph is arbitrary, the relative scaling between the two sectors is fixed by our construction and is quite consistent with intuition. It assigns greater weight to the extreme elements of the larger sector than to the extreme elements of the smaller sector. This is natural because we have asserted that we believe that the first element of the sector with fifty elements dominates forty-nine other components, whereas the first element of the sector with ten elements dominates only nine other assets.

4.4.3 Complete sort with long–short beliefs

As a modification of the above case, we imagine that the stocks are divided into two groups: a 'long' group that we believe will go up, and a 'short' group that we believe will go down. Within each group we additionally have ranking information. If ℓ is the number of long stocks, then these beliefs may be expressed as

$$r_1 \geq \cdots \geq r_\ell \geq 0 \geq r_{\ell+1} \geq \cdots \geq r_n$$

which is a total of $m = n$ beliefs. This includes the special cases $\ell = n$ when we believe all assets will have positive return, and $\ell = 0$ when we believe all will have negative return.

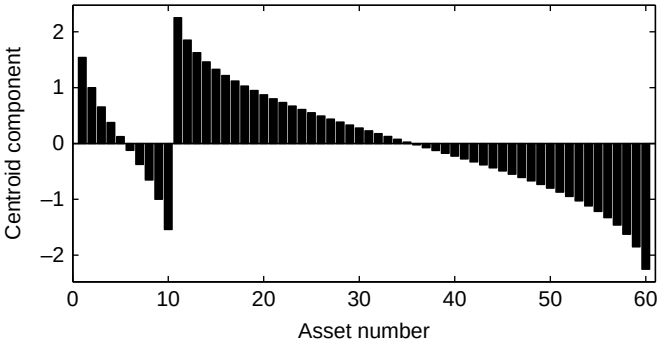


Figure 4.5 The centroid portfolio for two sectors of sizes $m_1 = 10$ and $m_2 = 50$.

To illustrate, let us take five assets, with the first two believed to have positive return, and the last three to have negative. Then $n = 5$, $\ell = 2$, and

$$D = \left(\begin{array}{cc|ccc} 1 & -1 & & & \\ & 1 & & & \\ \hline & & -1 & & \\ & & 1 & -1 & \\ & & & 1 & -1 \end{array}\right) \tag{4.6}$$

Orthogonal decomposition For a complete sort with long–short classification, there are *no* orthogonal directions since $m = n$. Every component of the return vector is relevant to our forecast. $R^{\perp} = \{0\}$, and $R = \mathbb{R}^n$.
Matrix structure The dual matrix is equation (4.6),

$$E = D^{-1} = \left(\begin{array}{cc|ccc} 1 & 1 & & & \\ & 1 & & & \\ \hline & & -1 & & \\ & & -1 & -1 & \\ & & -1 & -1 & -1 \end{array}\right)$$

Centroid profile Figure 4.6 shows the centroid vector with long/short constraints, for the case $n = 20$ and $\ell = 7$. This vector is not a simple linear transformation of the centroid vector without the zero constraint; its shape is more complicated.

4.4.4 Performance relative to index

We define an *index* to be a linear weighting of the assets

$$I = \mu_1 S_1 + \cdots + \mu_n S_n$$

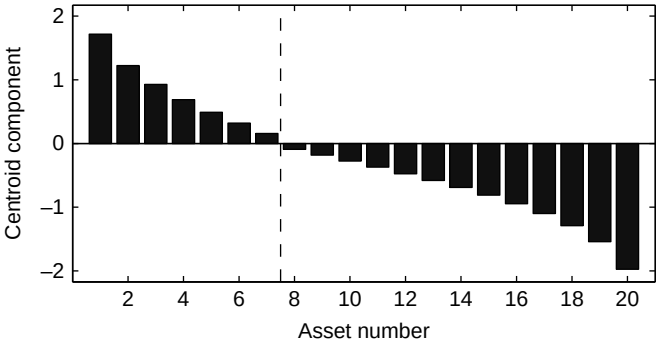


Figure 4.6 The centroid vector for a single sector of 20 assets, with sorting information plus the additional belief that the first seven will have positive return while the last 13 will have negative return.

with

$$\mu_j > 0 \text{ and } \mu_1 + \cdots + \mu_n = 1$$

We believe that the first ℓ stocks will *overperform* the index, and the last $n - \ell$ will *underperform*, with $0 < \ell < n$. Thus our beliefs are

$$r_j - (\mu_1 r_1 + \cdots + \mu_n r_n) \geq 0, \quad j = 1, \ldots, \ell$$
$$(\mu_1 r_1 + \cdots + \mu_n r_n) - r_j \geq 0, \quad j = \ell + 1, \ldots, n$$

and the belief matrix is

$$D = \begin{bmatrix} 1-\mu_1 & \cdots & -\mu_\ell & -\mu_{\ell+1} & \cdots & \mu_n \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ -\mu_1 & \cdots & 1-\mu_\ell & -\mu_{\ell+1} & \cdots & -\mu_n \\ \mu_1 & \cdots & \mu_\ell & \mu_{\ell+1}-1 & \cdots & \mu_n \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ \mu_1 & \cdots & \mu_\ell & \mu_{\ell+1} & \cdots & \mu_n - 1 \end{bmatrix}$$

Each of the belief vectors is orthogonal to $(1, \ldots, 1)^{\mathrm{T}}$. Thus the n belief vectors are in an $(n-1)$-dimensional subspace, and cannot be independents. But the cone $Q = \{Dr \geq 0\}$ has a non-empty interior $\{Dr > 0\}$; in fact it contains the vector $r = (1, \ldots, 1, -1, \ldots, -1)^{\mathrm{T}}$ for which

$$(Dr)_i = \begin{cases} 2(\mu_{\ell+1} + \cdots + \mu_n), & i = 1, \ldots, \ell \\ 2(\mu_1 + \cdots + \mu_\ell), & i = \ell + 1, n \end{cases}$$

Thus the theorems of Section 4.2 apply.

4.4.5 *Partial and overlapping information*

Below we review several different varieties of sorts that arise in practice.

Partial sorts

The centroid method works well for producing optimal portfolios from partial sorts – that is, from sorts that do not extend across an entire universe of stocks. The most natural way this arises in practice is in the case of a universe of stocks broken up into sectors. In this case a portfolio manager might have sorting criteria appropriate for stocks within a sector but which do not necessarily work for stocks across sectors.

Multiple sorts

In practice it is possible to have multiple sorting criteria. One might sort stocks according to their book-to-market ratio and size – for example, the logarithm of market capitalization. These characteristics provide two different sorts, but the resulting sorts are different and so it is impossible that they both be true. Nevertheless, both contain useful information that it would be suboptimal to discard.

Our portfolio optimization framework is valid in this case. Let Q_1 and Q_2 be the consistent cones under the two different criteria (e.g. Q_1 is the consistent cone for the book-to-market sort, and Q_2 is the consistent cone for the size sort). To apply our methodology, we have to supply a measure μ on the union of Q_1 and Q_2. We may do so by creating a density that assigns a probability p of finding an expected return in Q_1 and $1-p$ in Q_2. The centroid of the combined set under this measure is simply the weighted average of the two individual centroids. Using only the inequality information given by the sorts for Q_1 and Q_2 that have been specified, this is the only natural construction.

This formulation clearly applies to more than two non-overlapping weightings; we simply take the centroid of the combined set, which is an equal-weighted combination of the individual centroids. As an example, suppose that three orderings are given, and that two are close to each other. Then this algorithm will produce a centroid close to the centroids of the two close orderings.

Missing information

The above logic clearly indicates how to proceed when some information is considered unreliable. Suppose, for example, that it is believed that in the middle third of the asset list, the rankings have no significance. That is, the investor's belief is that all rankings within that subset are equally probable.

The extension of the above strategy says simply compute the superposition of the centroids of all the compatible orderings. The result of this is simply to average the centroid components within the uncertain index range.

4.5 Empirical tests

In this section we provide empirical examples of applications of optimization from ordering information using the linear, centroid, optimized linear and optimized centroid algorithm. Throughout this section we look to quantify the *absolute* and *relative* performance of the optimized centroid portfolios to portfolios constructed using the other methodologies. In particular, we would like to study the significance of incorporating covariance information into portfolio formation, especially as it pertains to improving performance relative to methods that omit this information. Since in practice many portfolio managers who use ordering information ignore covariance information, we believe that an important part of this work involves examining the extent to which using this method improves investment performance over existing methods. Put simply, we have proved the theoretical superiority of the centroid optimization method. Now we examine its practical significance.

We take two approaches to this, one based on studying an actual portfolio sort on real historical data, and the other based on simulations. In both approaches we create

backtests over a period of time, and for each time within the backtest period produce portfolios using four different construction methods from a single sort. The four methods we look at are the optimized and unoptimized versions of the linear and centroid methods, as described above.

To evaluate performance we examine the information ratio – that is, the annualized ratio of the sample mean and sample standard deviation of daily returns, of each time series derived from returns using the different construction methods. In our empirical work, we study reversal strategies, which, simply put, derive from the hypothesis that the magnitude of recent short horizon returns contains information about future returns over short horizons. Reversal strategies are easy to use to test portfolio construction methods using ordering information because in this case the ordering information is provided simply by recent returns. To be precise, we sort stocks within a universe of stocks based on the magnitude of past returns and then use the linear and centroid methods to construct portfolios and record their returns. The reversal work in this section demonstrates that the centroid optimization method provides a significant risk-adjusted performance boost over the unoptimized methods as well as a smaller, but significant, performance boost over the optimized linear method.

In our simulated work, we simulate markets using a stationary return-generating process of which we have complete knowledge. We then build portfolios based on the resulting asset sort and covariance information. We also look at the effect of creating portfolio sorts based on the correct asset ordering information after being re-arranged with a permutation. We study the relationship between the *variance* of the permutation – that is, the extent of the variance that the permutation introduces into the asset sort – and the resulting portfolio performance.

4.5.1 Portfolio formation methods and backtests

The empirical sections below show tests of the relative performance of different methodologies of forming portfolios from stock sorts. Both sections will take the same approach, though the first section will be a real empirical study while the second will be based on simulations.

To briefly review the definitions and results of Sections 4.2 and 4.3, all portfolio construction methods start with a list of portfolio constituents and a *portfolio sort*, which is an order or arrangement for the stocks in the portfolio $S_1, \ldots, S_n$ such that $r_1 \geq \cdots \geq r_n$, where r_i is the expected return for S_i. The portfolio formation methods are procedures for transforming a portfolio sort and (possibly) a covariance matrix into a portfolio. Here a portfolio is a list of dollar investments for each stock S_i in the portfolio, where the investment can be a positive number (representing a long position) or a negative number (representing a short position). In this instance, each portfolio is assumed to be held for a fixed time horizon at which point the portfolio is rebalanced (i.e. replaced) with a new portfolio. The return to such a portfolio over this fixed horizon is then defined as the sum of the products of the dollar investments multiplied by the stock returns over those fixed horizons. This is equivalent to looking at the dollar profit and loss to the portfolio over that period of time divided by the gross market value of the portfolio. We shall compare the relative performance of the four basic approaches for portfolio formation examined in this chapter: *linear, centroid, optimized linear* and *optimized centroid*.

In the linear and the centroid portfolios, the weights are specified directly, without use of the covariance matrix. The linear portfolio is the portfolio that forms a dollar-neutral portfolio with linearly decreasing weights, assigning the highest weight to the highest-ranked stock and the lowest weight to the lowest-ranked stock. Variants of this portfolio are used in practice quite often in the asset management community. The (unoptimized) centroid portfolio is the portfolio formed by assigning a weight proportional to that of the centroid vector to each stock according to the rank of the stock. Intuitively, this portfolio is similar to the linear portfolio but with the *tails* overweighted; that is, the highest- and lowest-ranked stocks receive proportionately more weight in this portfolio than the linear one.

The two 'optimized' construction methods make use of the covariance matrix and are extremal in the sense described in Section 4.2. Roughly speaking, they construct mean-variance optimal portfolios as if the expected returns for the asset universes have a respectively linear (for the optimized linear) or centroid (for the optimized centroid) profile with respect to their rankings.

For the purposes of this section, a backtest is a method for testing the joint procedure of forming a portfolio sort and constructing a portfolio using a particular construction method. A backtest tests the *investment performance* of a particular portfolio sorting procedure (for example, the reversal sort) combined with a particular portfolio construction method. The basic premise is that a superior portfolio sort combined with a superior portfolio construction method will produce a superior risk-adjusted return. A backtest then produces a time series of investment returns based on repeated application of a consistent portfolio construction method to a particular portfolio sorting procedure. Performance is measured by the information ratio which measures the return per unit of risk as the annualized ratio of the average return of the portfolio to the standard deviation of that return.

In all that follows we will produce backtests for a single portfolio sorting method while forming all four of the above portfolios (linear, centroid, optimized linear, optimized centroid) and compare information coefficients.

A backtest has several major components:

- *Periodicity*. This is the time frequency over which portfolio returns are produced. A daily backtest produces portfolio returns for each day.
- *Length*. This is the number of time steps (where each time step's length is determined by the periodicity) of the backtest.
- *Start date*. This is the first date on which a return is produced for the backtest.
- *End date*. This is the last date on which a return is produced for the backtest.
- *Portfolio formation dates*. These are dates on which portfolios are formed within the backtest.

In our backtests below we will use the following basic procedure for each portfolio formation date between a specified start date and end date:

- Determine portfolio sort
- Compute covariance matrix
- Form the four portfolios: linear, centroid, optimized linear, optimized centroid
- Determine return of each portfolio held from formation date to next formation date.

In each step of the above procedure, we take care to avoid look-ahead bias in both our covariance computations and the production of our sorts.

4.5.2 Reversal strategies

In this empirical example we look at *reversal strategies* that seek to buy stocks whose prices appear to have moved too low or too high due to *liquidity* and not due to *fundamental* reasons. Effectively, the strategy seeks to buy stocks at a discount and sell stocks at a premium to fair prices in order to satisfy a counterparty's requirements for immediacy.

The theoretical underpinnings for reversal strategies and empirical evidence for them are discussed in Campbell *et al.* (1993). These authors looked specifically at stocks with large price movements accompanied by large trading volume and measured observed serial correlation around these events. We call the hypothesis that stock prices reverse in the manner just described the *reversal hypothesis*, and take as given that trading strategies exploiting these strategies will produce positive expected returns.

We build a simple portfolio sort to exploit the tendency for stocks to reverse. We produce the sort from the magnitude of recent past returns and sort stocks from highest to lowest according to their negative. Stocks whose past returns are most negative are therefore deemed most favourable, while stocks whose returns are most positive are viewed least favourably. This provides us with a straightforward method for demonstrating the effectiveness of the portfolio construction methods in this chapter. We do not explicitly form expected returns from the information about past returns, we only sort stocks according to this information and construct portfolios accordingly.

Data, portfolio formation and reversal factor

Since we aim to produce an overall picture of the relative performance of the four portfolio formation methods, we will look at the reversal strategy across a range of possible variations of its basic parameters. Below is a list of the key variants in the reversal strategy tests.

- The *portfolio size* N is the number of stocks in the portfolio on any given date. For all of our studies, the number of stocks in the portfolio remains constant over the test period. We study portfolios of sizes 25, 50, 100, 150, 200, 250, 300 and 500.
- The *reversal period* K is the number of days of return data to include in the computation of the reversal factor. We study strategies with reversal periods 5, 10, 15, 20 and 25. For example, a reversal period of 5 means that we sort stocks based on past returns over 5-day periods.
- The *reversal lag* L is the number of days prior to a given portfolio formation date that the reversal factor is computed over. We study strategies with reversal lags 0 and 1.

We use the Center for Research in Security Prices (CRSP) database of daily US stock prices from NYSE, Amex and Nasdaq. The *return* of a stock on day t is the CRSP *total return* of the stock: the arithmetic return of the price from day $t-1$ to day t plus the effect of dividends and other income that would accrue to the investor over that period.

For each reversal strategy variant above the following procedure is undertaken. We start with the CRSP database of returns from 19 January 1990, to 31 December 2002.

On the first date of this period we choose a *universe* consisting of the 1000 largest stocks sorted by capitalization for which there also exist at least 1000 valid days of data prior to the start date. On each subsequent date if a stock drops out of the universe (e.g. if the stock is delisted, merges or goes bankrupt), we replace it with a new stock that (a) is not already in our universe, (b) has at least 1000 valid days of data through that date, and (c) is the largest possible stock that meets the criteria (a) and (b). After this universe is created, reversal strategy parameters are selected: values are chosen for portfolio size N, reversal lag L and reversal period K. From these data a backtest is conducted as follows:

- For each date t in the backtest period we form a *portfolio list* for date t, by criteria (a), (b) and (c) above. The portfolio list is the list of candidate stocks for the portfolio on that date.
- The N *constituents* of the portfolio formed on date $t-1$ are chosen randomly from the portfolio list.
- The sort parameter on date $t-1$ is the *negative* of the cumulative total return from date $t-L-K$ to $t-L-1$. The stocks are rearranged in decreasing order by this variable. Thus, stock with index i is always the stock whose performance is expected to be ith from the top, but this will be a different stock on different dates.
- For a portfolio formed on date $t-1$ we compute a *covariance matrix* from a rectangular matrix of data consisting of the N columns and $2N$ rows of data from days up to but not including day t. By construction, all elements of this rectangular grid contain a valid total return for the portfolio constituents at time $t-1$.
- For each date t in the backtest period we form linear, centroid, optimized linear and optimized centroid portfolios on date $t-1$ using a portfolio sort and covariance matrix. The portfolios are normalized to have unit *ex ante* risk as measured by the estimated covariance matrix. They are not constrained to be market neutral.
- The portfolio formed on day $t-1$ is held from $t-1$ to t. Its return over this period is calculated as the sum of the product of the return of each stock in the portfolio on day $t-1$ multiplied by the portfolio holding in that stock on day $t-1$. The portfolio is assumed to be rebalanced at the close of day t (i.e. traded at a price equal to the closing price on day t and then held until day $t+1$). In this way the entire backtest is run.

Comments

Our method for producing the active universe from which to form portfolios was designed to create a universe which had no look-ahead or survivorship biases present, and which was compact enough to allow us to use simple methods for computing the covariance matrices. We make no attempt to measure transaction costs or to form realistic trading strategies to minimize turnover. (It is anecdotally known that even trading strategies with reported information ratios as high as 5 before transaction costs and commissions still do not return a profit in actual trading.)

While our strategies do not suffer from look-ahead or survivorship bias, they are not necessarily realistic. For example, when the reversal lag is set to zero, the reversal factor is literally formed from returns data including the same day's closing prices as is supposedly captured in the rebalance. Therefore, a reversal lag of zero is technically impossible, but actually the limit as time tends to zero of a procedure which is technically

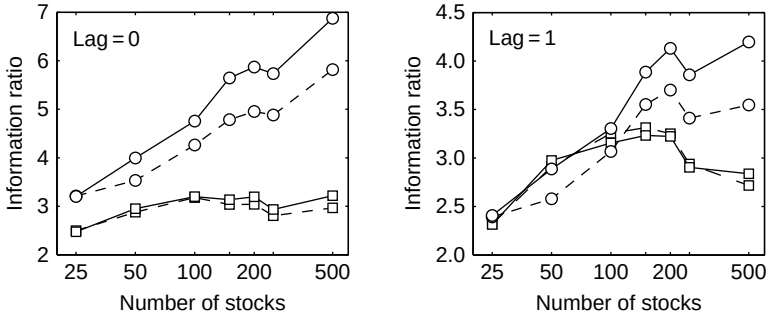
possible once transaction costs are taken into account. Nevertheless, the procedure does not incorporate data in any way that could not be (at least theoretically) known prior to portfolio formation. In addition, while a reversal lag of zero is impossible, a reversal lag of one is unrealistically long since no realistic trading strategy would form a factor and wait an entire day before trading.

The key aim of this example is to examine the *relative* performance of a strategy based on centroid optimization versus a linear weighting scheme. We are not attempting to test either the availability of excess returns or the validity of a certain hypothesis about a certain pricing anomaly. In fact, we take as given that reversal hypothesis holds. It is obvious that building a portfolio from reversal information will produce positive returns if the reversal hypothesis holds. It also *seems* obvious that by using only information about the relative strength of the expected reversal and not information about the covariance structure of the stocks in the universe, one cannot expect to construct a portfolio with maximum information ratio. What is not obvious, however, is the degree to which the covariance structure can improve the information ratio. This is the empirical piece of information that we are attempting to calibrate.

Backtest results for reversal strategies

Figure 4.7 shows a concise summary of the results. For a reversal period of five days, we show the information ratios (mean return divided by its standard deviation) for the four strategies described above, for lag of zero and lag of one day. Table 4.1 shows the full set of results.

The information ratios of the unoptimized portfolios are consistently around three or less. Use of the covariance matrix dramatically improves the results: both the optimized linear and the optimized centroid algorithms achieve information ratios that are



From bottom to top the curves represent the unoptimized linear, unoptimized centroid, optimized linear and optimized centroid constructions; the linear portfolios are drawn with dashed lines and the optimized portfolios are drawn with round markers. The left panel shows reversal lag of zero, meaning that information is used instantly; the right panel shows a lag of one day. For large portfolios, the optimized centroid gives more than a twofold improvement over the unoptimized linear, and is substantially better than the optimized linear.

Figure 4.7 Mean realized information ratios of the four algorithms described in this chapter, using a reversal strategy with reversal period of five days.

Table 4.1 Information ratios for the four strategies considered in this paper, for all combinations of input parameters

Number of stocks	Reversal period (days)									
	5		10		15		20		25	
25	2.50	2.47	2.36	2.40	1.72	1.75	1.42	1.45	1.59	1.61
	3.21	3.20	2.37	2.50	1.84	1.95	1.69	1.93	1.63	1.79
50	2.88	2.95	2.92	3.10	2.39	2.52	2.07	2.16	2.06	2.14
	3.53	3.99	3.26	3.63	3.03	3.35	2.93	3.30	2.79	3.07
100	3.18	3.20	2.98	3.12	2.46	2.61	2.09	2.17	2.17	2.19
	4.26	4.76	3.65	4.09	3.54	3.95	3.19	3.73	3.10	3.43
150	3.04	3.14	2.83	2.99	2.57	2.69	2.21	2.33	2.29	2.34
	4.79	5.65	4.15	4.82	4.13	4.80	3.69	4.44	3.53	3.97
250	2.80	2.93	2.43	2.64	2.16	2.33	1.94	2.10	2.06	2.20
	4.88	5.73	3.67	4.47	3.61	4.27	3.39	4.12	3.13	3.68
500	2.97	3.22	2.40	2.72	2.11	2.37	1.91	2.19	1.93	2.16
	5.82	6.88	4.33	5.38	4.31	5.25	4.44	5.40	4.31	5.10
25	2.32	2.32	2.10	2.15	1.61	1.61	1.20	1.24	1.46	1.46
	2.39	2.41	1.84	1.86	1.35	1.40	1.25	1.43	1.35	1.48
50	2.91	2.97	2.80	2.96	2.29	2.37	1.85	1.90	1.93	1.97
	2.58	2.89	2.52	2.85	2.46	2.65	2.32	2.54	2.30	2.48
100	3.25	3.15	2.77	2.84	2.29	2.38	1.81	1.85	1.95	1.94
	3.07	3.30	2.47	2.83	2.62	3.02	2.34	2.87	2.38	2.68
150	3.31	3.23	2.68	2.73	2.38	2.44	1.89	1.98	2.00	2.04
	3.55	3.89	3.07	3.48	3.24	3.69	2.72	3.28	2.75	3.05
250	2.94	2.90	2.22	2.32	1.87	1.97	1.59	1.69	1.68	1.80
	3.41	3.86	2.41	2.90	2.35	2.88	2.26	2.86	2.10	2.54
500	2.72	2.84	1.95	2.15	1.60	1.81	1.36	1.59	1.37	1.60
	3.55	4.20	2.27	2.95	2.47	3.19	2.74	3.46	2.71	3.30

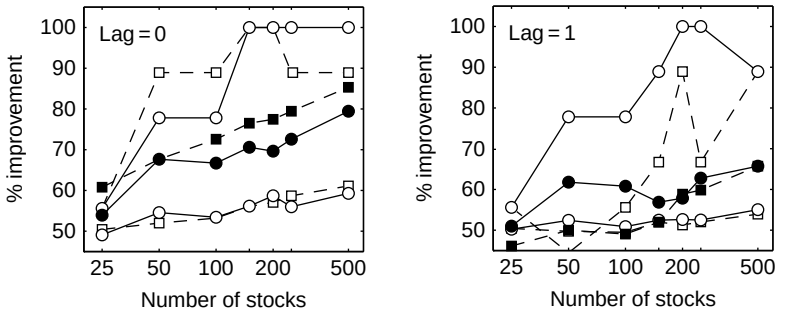
Upper panel is zero lag, lower panel is lag of one day. Within each box, the left column is based on the linear portfolio, the right on the centroid; the upper row is the unoptimized portfolios and the lower row is the optimized portfolios.

always better than either of the unoptimized versions, and the improvement increases logarithmically with portfolio size.

Most significantly, for all portfolio sizes the optimized centroid algorithm performs substantially better than the optimized linear algorithm, although the two unoptimized portfolios are essentially equivalent. This supports the main argument of this chapter, which is that the centroid construction is the best way to combine covariance information with sorts. The results shown here cannot be achieved for any portfolio construction algorithm based on linear profiles.

The improvement is weaker for a lag of 1 day, but still substantial. For longer reversal periods, the results are similar though less dramatic.

To reinforce this point, in Figure 4.8 we show the percentage of non-overlapping days, months and years for which our optimized centroid algorithm realizes a higher return than its two competitors: the unoptimized linear algorithm and the optimized linear algorithm. For example, in the monthly returns (heavy lines) the optimized centroid realizes a better return over 80% of the time for large portfolios, with zero lag.



From bottom to top, the curves show 1-day, 20-day (heavy lines) and 250-day non-overlapping periods.

Figure 4.8 Percentage of days, months, and years on which the optimized centroid portfolio realizes a higher return than the unoptimized linear portfolio (dashed lines, square markers) and than the optimized linear portfolio (solid lines, round markers).

Note that this is *not* a direct test of the reasoning underlying our centroid algorithm. In Section 4.3 we constructed the centroid vector to maximize the fraction of realizations of the *expected* return vector for which it would be optimal. In this test we measure the *actual* returns, which depend additionally on the covariance structure and any possible non-Gaussian properties. Our intention has been to provide a realistic and financially meaningful test of the algorithm in practice.

4.5.3 Simulation results

We now seek to explore our portfolio construction methods in an environment where we have complete control and knowledge of all the return generating processes and portfolio sorts. This section has two aims. This first is to test the limits of the method and understand in numerical terms what the limits of the method are when perfect knowledge of portfolio sorts is available. The second is to characterize the impact of information degradation on our methods. We do this by turning perfect knowledge of portfolio sorts into less-than-perfect knowledge of portfolio sorts by introducing the effect of a permutation on the precise expected return ordering and studying the ensuing breakdown of portfolio performance.

We run backtests using the portfolio construction methods described in this chapter and using the same structure as in the reversal tests above. In this case, however, we simulate stock returns and have perfect knowledge of the key elements necessary to estimate our models. That is, we retain perfect knowledge of the covariance matrix and order of expected returns. We do this in a variety of scenarios across different portfolio sizes (from 25 to 500 stocks) and different volatility structures (from very low levels to very high levels of cross-sectional volatility). For each scenario we run multiple iterations and record the average performance across iterations to ensure that the results are representative of the average expected performance levels. Most importantly, we study the *degradation* of performance, both absolute (in terms of information ratios) and relative (in terms of the improvement of optimized methods to non-optimized methods), as we degrade

information. We introduce a measure of information distortion generated by examining the amount of variance introduced into the system by permuting the indices of the correct order and relate this to the correlation coefficient of a Gaussian copula.

Simulated stock returns

In order to simulate backtests of our results, we simulate stock returns for a universe of stocks and then use these return histories as in Section 4.5.1. Presently, we discuss in detail the parameters that define our simulated stock histories.

As we are mainly interested in bounding the overall expected levels of performance of the portfolios that we construct from sorting data, we focus on variables that we believe *ex ante* provide the greatest degree of variability in overall portfolio performance. In our view these are cross-sectional volatility and expected return spread, and number of stocks in the portfolio. Cross-sectional volatility refers, roughly speaking, to the variability of volatility at a point in time in the cross-section of stocks in the universe. Return spread refers to the differential in expected return spread in the cross-section of stocks.

A critical other variable that will clearly determine the success of our methods is the extent of one's knowledge of the order of expected returns. In the main text of this chapter we assumed perfect knowledge, but in practice perfect knowledge is impossible to come by. Therefore we turn our sights to performance in the presence of information that is correlated to, but not identical to, the precise order of expected returns.

We simulate the returns of a portfolio by assuming that variation among its constituents is generated by a system consisting of a common factor and an idiosyncratic component. We assume that we have a portfolio with N stocks $S_1, \ldots, S_N$ whose expected returns $r_1, \ldots, r_N$ are in descending order – that is, they satisfy the inequalities $r_1 \leq \cdots \leq r_N$.

In our simulations, the *realized* return r_{it} of stock i at time t is generated by the factor equation

$$r_{it} = F_t + \varepsilon_{it} + \mu_i \tag{4.7}$$

where F_t is regarded as a 'factor return' at time t. We assume that the F_t are independent, identically distributed normal random variables. Similarly, ε_{it} is 'idiosyncratic risk' and for a fixed i, the ε_{it} are independent, identically distributed normal random variables. We have that each ε_{it} is mean zero and with variance set to the number $\sigma_i^2/2$; likewise F is a normally distributed random variable with mean zero and variance equal to the average variance of the ε_i, that is:

$$\sigma^2(F) = \frac{1}{N}\sum_i \frac{\sigma_i^2}{2}$$

The σ_iS are set to be equally spaced in log space and sorted so that

$$\sigma_1 < \sigma_2 < \cdots < \sigma_N$$

That is, if $\sigma_{\min}$ and $\sigma_{\max}$ are respectively associated with the minimum and maximum idiosyncratic volatility, then if $\sigma_1 = \sigma_{\min}$ and $\sigma_{\max}$ we have that

$$\log \sigma_1, \log \sigma_2 \ldots, \log \sigma_N$$

are equally spaced. We do this so that in the space of stocks the occurrence of high-volatility stocks is less frequent then that of low-volatility stocks. Also, we specify the *distance* in log-space between the minimum and maximum volatility, δ, and call this the *volatility dispersion*. It is defined by

$$\sigma_{\max} = \delta\sigma_{\min}$$

Finally, the variable μ_i is a constant for each simulation that defines the cross-sectional expected return spread. It is sorted so that

$$\mu_1 \leq \mu_2 \leq \cdots \leq \mu_N$$

and the vector $(\mu_1, \ldots, \mu_N)$ is generated as the ascending sort of N i.i.d. draws from a normal distribution with mean and variance both equal to $0.6/16 \cdot \sigma^2(F)$. This means that the average daily Sharpe ratio in the cross-section of stocks is approximately 0.6/16 so that the annualized Sharpe ratio of each stock is approximately 0.6.

The volatilities of the stocks in the cross-section are calibrated as follows. For each simulation run we choose a volatility dispersion δ and build equal log-spaced volatilities, σ_i, with $\sigma_{\min} = 0.005/\sqrt{(2)}$ and $\sigma_{\max} = \delta \cdot \sigma_{\min}$.

To relate this characterization of volatility dispersion to US markets, we briefly examine recent market behaviour. Volatility dispersion is somewhat difficult to measure in practice due to outliers in the cross-section of volatility within the US market. But to give a sense, in a randomly selected sample of 500 of the largest US stocks we have the following observations. We compute volatility by computing the annualized standard deviations of past 250-day returns. For a given day we then compute the cross-sectional standard deviation of volatility, and trim those volatilities which exceed three standard deviations above or below the mean. Using such a measure, the dispersion of volatility has ranged from 4.31 to 16.7 in our data set.

Permutations and information distortion

In practice, we do not expect portfolio managers to know the exact order of the expected returns of a list of stocks. Rather, we expect a proposed ordering to be *correlated with* the true ordering of the expected returns. In this section we provide a measurement of the difference between the exact ordering of expected returns and a permutation of that ordering. To describe this, we begin with a list of stocks

$$S_1, S_2, \ldots, S_N$$

whose true expected returns satisfy

$$r_1 \leq r_2 \leq \cdots \leq r_N$$

A permutation π of the list is a mapping

$$\pi : \{S_1, \ldots, S_N\} \rightarrowtail \{S_{\pi(1)}, \ldots, S_{\pi(N)}\}$$

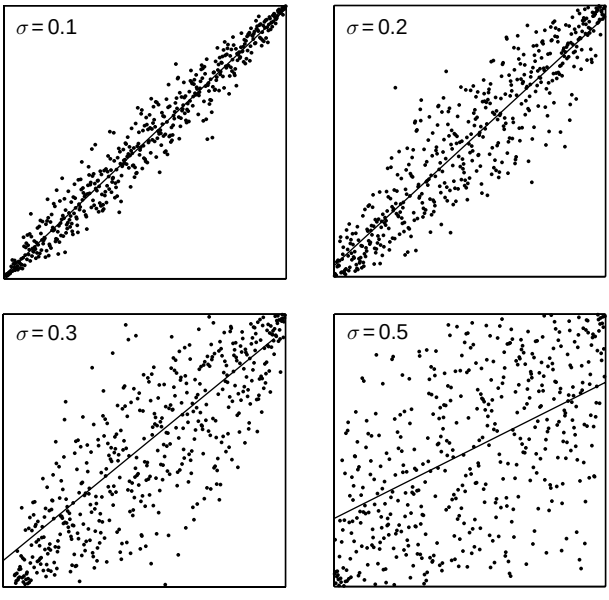
representing the relationship between the true ordering and our partially erroneous information. The *minimally distorting permutation* $\pi_{\min}$ is the identity map $(S_1, \ldots, S_N) \rightarrowtail (S_1, \ldots, S_N)$, representing perfect information. The *maximally distorting permutation* $\pi_{\max}$ is the permutation that completely reverses the order of the indices: $(S_1, \ldots, S_N) \rightarrowtail (S_N, \ldots, S_1)$ representing information that is perfectly wrong. We define the *distance* σ of a permutation π to measure position between these two extremes:

$$\sigma(\pi) = \sqrt{\frac{\Sigma(\pi(i) - i)^2}{\Sigma(\pi_{\max}(i) - i)^2}} = \sqrt{\frac{1}{2}(1 - b)}$$

where b is the coefficient in the linear regression to the points $(i, \pi(i))$.

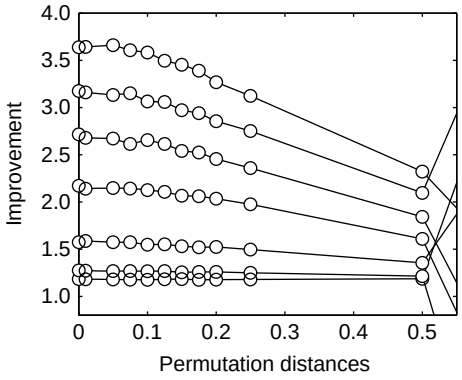
Thus, permutation distance $\sigma(\pi)$ is a number between zero and one, that measures the degradation of our information about the order of the expected returns. It is a general measure of information *loss*. Figure 4.9 provides a graphical representation of permutation distances. As permutation distance increases from zero, the quality of our information about the sort decreases; maximal uncertainty, or complete randomness, is obtained at $\sigma = 1/\sqrt{2}$.

For any value of ε there are many permutations whose distance is approximately ε. Naturally, for a given ε, there is a finite space of permutations of N indices with distance ε which have varying characteristics. As a consequence, in our simulations below where we study the impact of permutation of order on investment performance, we are careful



The points have coordinates $(i, \pi(i))$, where i is a permutation index and $\pi(i)$ is its destination under the permutation; the line is the linear regression of slope $1 - 2\sigma^2$. A distance of 0.71 is completely random.

Figure 4.9 Permutations with specified distances, for $N = 500$.



The horizontal axis is permutation distance as defined in the text. From bottom to top, the curves have volatility dispersion 1, 2, 4, 8, 12, 16, 20. The improvement is quite dramatic when substantial dispersion is present, and degrades on very slowly.

Figure 4.10 Mean improvement in Sharpe ratio of the risk-optimized centroid algorithm relative to the linear portfolio weighting, for simulated portfolios of 250 assets.

to compute multiple iterations of simulations for given permutation distances, sampling randomly from the space of correlations.

Simulation results

For each scenario the following parameters apply:

- *Portfolio size*. The portfolio size describes the number of stocks N in the portfolio. In our simulations, the portfolio sizes vary from 25 to 500 stocks.
- *Factor structure*. Factor structure describes the structure of the return-generating process for the stocks in the simulated portfolio, and is given by equation (4.7).
- *Volatility dispersion*. As described above. In our simulations, volatility dispersion ranges from 1 to 20.
- *Permutation distance*. We generate 10 permutations with distances equal to 0, 0.01, 0.05, 0.075, 0.1, 0.125, 0.15, 0.175, 0.2, 0.25, 0.5 and 0.7071.
- *Simulation length*. Simulation length describes the number of time steps in each simulation. In this backtest the simulation length is always set to 2000, meaning that each simulation simulates a backtest of 2000 time steps.
- *Iterations*. The number of iterations computed for each set of parameters over which to compute statistics. For all simulation parameter sets in this chapter, we compute 50 iterations.

Our simulated results provide insight into the possible improvements that our methodologies provide. All of the basic relationships hold. Information ratios increase with volatility dispersion, and numbers of stocks in our portfolios increase. Also, holding other factors constant, information ratios decrease as permutation distance increases. This indicates, as expected, that as the quality of a portfolio manager's sort decreases, so does investment

Table 4.2 Information ratios for backtests run on simulated markets generated with a one factor model with volatility dispersions and permutation distance (information degradation) as shown

Vol disp	Permutation distance (σ)									
	0		0.1		0.2		0.5		0.7	
1	18.1	18.6	18.0	18.3	16.9	17.2	9.5	9.6	0.0	0.0
	21.0	21.4	20.7	21.1	19.5	19.9	11.1	11.2	0.0	0.0
2	17.3	17.5	17.1	17.3	16.2	16.3	9.1	9.2	−0.1	−0.1
	21.4	22.0	21.0	21.6	19.9	20.3	10.8	11.0	0.1	0.2
4	15.4	15.2	15.3	15.1	14.5	14.4	8.5	8.6	0.0	0.0
	23.4	24.3	22.9	23.6	21.5	22.1	11.3	11.5	−0.2	−0.3
8	13.3	12.8	13.1	12.6	12.7	12.2	7.8	7.7	0.0	0.0
	27.2	28.8	26.4	27.8	24.7	25.8	12.3	12.5	0.5	0.5
16	11.6	11.0	11.5	10.9	11.2	10.6	6.9	6.8	0.0	−0.1
	34.3	36.8	33.0	35.1	30.3	31.9	14.1	14.4	−0.1	0.0
20	11.1	10.4	10.9	10.3	10.6	9.9	6.8	6.7	0.0	0.0
	37.4	40.4	36.4	39.2	32.6	34.5	15.3	15.7	−0.8	−0.8

Each cell of this table represents the mean information ratios for the four portfolio construction algorithms described in this paper laid out as in Table 4.1. All portfolios have 500 stocks and all backtests are run over a period of 2000 days. Stocks are calibrated so that they have on average an annualized information ratio of 0.6, assuming each time-step is one day and each year has 256 days. Information ratios are annualized assuming that each time step is one day and each year has 256 days.

performance. Of course, we also see that the algorithm is highly robust to information loss. On examining Table 4.2 we see that even for high degrees of information loss, the strategies still provide significant return on risk.

An important observation which can also be seen in Table 4.2 is that, holding other factors constant, as permutation distance increases the extent of the improvement between the optimized centroid and optimized linear algorithm narrows. For example, for 500 stocks, a volatility dispersion of 20 and with zero permutation distance (the perfect knowledge scenario), the optimized linear algorithm provides an average information ratio (over 50 trials) of 37.4 while the optimized centroid provides an information ratio of 40.4, an improvement of over 8%. For a permutation distance of 0.5, the optimized linear algorithm provides an average information ratio of 15.3 while the optimized centroid provides an information ratio of 15.7, a spread of only 2.6%.

The above-described pattern of contracting optimized centroid performance in the presence of information degradation is present throughout the table. In an indirect way it confirms one of the central arguments of this chapter, which is that the optimized centroid outperforms not only the unoptimized algorithms but also other *extremal* optimized algorithms. An intuitive explanation for this may be found by recalling the nature of the optimized centroid algorithm. In our simulations, roughly speaking subject to a risk budget, the optimized centroid algorithm maximizes exposure to difference assets.

Difference assets are assets that express a single belief, such as, 'Stock one is better than stock two'. Such a belief is expressed by means of a difference asset by maximizing exposure to such an asset. For the preceding example, the belief is that the asset would be $D = S_1 - S_2$. Now, what does 'perfect information' concerning the order of expected

returns mean in the context of difference assets? The answer is clear: it implies that every difference asset has a positive expected return. Now, when we introduce permutations, that is information degradation, into the picture, what happens is that we switch some of the difference assets from having positive to having negative expected returns. The upshot of this is that the algorithm which maximizes exposure to the difference assets should have the most rapid degradation of relative performance due to the introduction of permutations. This naturally suggests a possible avenue of further research. Is there a robust version of the centroid algorithm which is better able to deal with information degradation in assuming that a certain percentage of the difference assets might swap from positive to negative expected returns? And, would such an algorithm outperform the centroid algorithm in real-life scenarios?

4.6 Conclusion

This chapter began with a simple question: What is the best way to form an investment portfolio given a list of assets, an ordering of relative preferences on those assets, and a covariance matrix? A large part of the motivation came from the observation that existing methods were very *ad hoc* and had no way of incorporating volatility and correlation information. These methods were therefore incompatible with the main stream of portfolio methods going back to Markowitz, who emphasized the importance of incorporating risk into the construction of the optimal portfolio.

In the course of developing this solution, we were led to develop a very robust and powerful framework for thinking about portfolio optimization problems. This framework includes 'classic' portfolio theory as a special case and provides a natural generalization to a broad class of ordering information. It also includes more modern constructions, such as robust optimization, as we now discuss.

To summarize, our formulation has three ingredients:

1. *Ordering information* which gives rise to a *cone of consistent returns*. This is a set in which the true expected return vector is believed to lie. In the examples considered in this chapter, this cone is always constructed as the intersection of half-spaces corresponding to a finite list of homogeneous inequality beliefs. But more generally, we may specify any convex cone, with curved edges or other more complicated geometrical structure; our construction can in principle be carried out for any such set.
2. A *probability density* within the belief cone: a measure that specifies our belief about the relative probability of the actual expected return vector being any particular location within the belief cone. In this chapter we have considered only the case in which this density has radial symmetry, but the framework also allows more general densities.
3. A *constraint set* in which the portfolio is constrained to lie. In the most important application, this constraint set is determined by a total risk limit given by a covariance matrix, but it may be any convex set. In particular, it may include short sales constraints or position limits; all the standard techniques of constrained optimization may be brought into play.

Empirical tests show that the resulting portfolios are substantially better than the ones given by the *ad hoc* formulations.

In classic Markowitz theory, a single expected return vector is given. In our formulation, that single vector generates a half-space of possible expected return vectors that have non-negative inner product with the given vector. For any given constraint set, our construction of the efficient set and then the optimal portfolio gives the identical result to the Markowitz theory. If further inequality constraints are then added, we naturally incorporate those beliefs.

In robust optimization, it is recognized that the actual expected return vector may not be exactly equal to the single given vector. In effect, a probability density is introduced that is centred on the given vector, and various minimax techniques are used to generate optimal portfolios. In our framework, the density would be modelled directly, and it could additionally be constrained by inequality relationships.

Our construction provides a rich and flexible framework for exploring the nature of optimal portfolios. In future work, we plan to consider some of these extensions and applications.

Appendix A: Proofs of theorems

In this Appendix, we prove Theorems 1 and 2 of Section 4.2.4, and justify the need for the interior in Theorem 2.

Proof of Theorem 1

First, suppose that $b \in \overline{Q}$ is a strict normal. If there is a $v \in \mathcal{M}$ with $v \succ w$, then in particular $v \succeq w$, that is, $(v-w)^{\mathrm{T}} r \geq 0$ for all $r \in \overline{Q}$. But this contradicts the hypothesis.

Second, suppose that $b \in \overline{Q}^{\circ}$ is a normal – that is, that $b^{\mathrm{T}}(v-w) \leq 0$ for all $v \in \mathcal{M}$. Suppose there is a $v \in \mathcal{M}$ with $v \succeq w$ (that is, $(v-w)^{\mathrm{T}} r \geq 0$ for all $r \in \overline{Q}$); then $(v-w)^{\mathrm{T}} b = 0$. Since $b \in \overline{Q}^{\circ}$, for any $s \in R$, $b + \varepsilon s \in \overline{Q}$ for ε small. Then $(v-w)^{\mathrm{T}}(b+\varepsilon s) \geq 0$, which implies that $(v-w)^{\mathrm{T}} s = 0$ for all $s \in R$, so $v - w \in R^{\perp}$. But then it is impossible that $(v-w)^{\mathrm{T}} r > 0$ for any $r \in \overline{Q}$, so $v \not\succ w$.

Simple examples show that the strictness conditions are necessary.

In proving Theorem 2, we must make use of the set

$$
\begin{aligned}
K &= \{w \in \mathbb{R}^n | w \succ 0\} \\
&= \{w | w^{\mathrm{T}} r \geq 0 \text{ for all } r \in \overline{Q} \text{ and } w^{\mathrm{T}} r > 0 \text{ for at least one } r \in \overline{Q}\} \\
&= \{w | w_0^{\mathrm{T}} r \geq 0 \text{ for all } r \in Q \text{ and } w_0^{\mathrm{T}} r > 0 \text{ for at least one } r \in Q\}
\end{aligned}
$$

By the last representation, we have $K = K_0 \oplus R^{\perp}$ with $K_0 \subset R$.

Lemma: K is convex, and $K \cup \{0\}$ is a convex cone.

Proof: For $w_1,\ w_2 \in K$ and $\alpha,\ \beta \geq 0$, we must show $\overline{w} = \alpha w_1 + \beta w_2 \in K$ for α and β not both zero. Clearly $\overline{w}^{\mathrm{T}} r \geq 0$ for $r \in \overline{Q}$. And letting $r_i \in \overline{Q}$ be such that $w_i^{\mathrm{T}} r_i > 0$ and setting $\overline{r} = \alpha r_1 + \beta r_2$ ($\overline{Q}$ is a convex cone), we have $\overline{w}^{\mathrm{T}} \overline{r} = \alpha^2 w_1^{\mathrm{T}} r_1 + \beta^2 w_2{}^{\mathrm{T}} r_2 + \alpha\beta(w_1^{\mathrm{T}}\ r_2 + w_2^{\mathrm{T}} r_1) > 0$.

Lemma: If $\overline{Q}^\circ$ is non-empty, then $\overline{Q}^* \subset K \cup R^\perp$.

Proof: Take any $w \in \overline{Q}^*$, so $w^\mathrm{T} r \geq 0$ for all $r \in \overline{Q}$. Choose $r_0 \in \overline{Q}^\circ$; then $r_0 + \varepsilon s \in \overline{Q}$ for all $s \in R$ and for all ε small enough. If $w \notin R^\perp$, then there are s so $w^\mathrm{T} s \neq 0$ and hence if $w^\mathrm{T} r_0 = 0$ there would be $r \in \overline{Q}$ with $w^\mathrm{T} r < 0$. Thus we must have $w^\mathrm{T} r_0 > 0$ and $w \in K$.

Lemma: If $\overline{Q}^\circ$ is non-empty, then $\overline{Q}^* \subset \mathrm{cl}(K)$, where $\mathrm{cl}(\cdot)$ is closure.

Proof: We need to show that $R^\perp \subset \mathrm{cl}(K)$. But the previous lemma but one showed that $0 \in \mathrm{cl}(K)$, and the result follows from $K = K_0 \oplus R^\perp$

The following facts are more or less standard; they are either proved in Boyd and Vandenberghe (2004) or quite simple:

If $A \subset \mathbb{R}^n$ is a closed convex cone, then $(A^*)^* = A$.

If $A \subset B$, then $B^* \subset A^*$.

$\mathrm{cl}(A^*)^* = A$.

Proof of Theorem 2

Since w is efficient, the convex sets $w + K$ and $\mathcal{M}$ are disjoint. By the supporting hyperplane theorem, $\mathcal{M}$ has a supporting hyperplane at w whose normal b has $b^\mathrm{T}(v - w) \geq 0$ for all $v - w \in K$. That is, $b \in K^*$ and we need $b \in \overline{Q}$: we must show $K^* \subset \overline{Q}$.

This follows from the lemmas: $K^* = \mathrm{cl}(K)^* \subset (\overline{Q}^*)^* = \overline{Q}$.

This theorem is also true if $\mathcal{M}$ is the boundary of a convex set, since it relies only on the existence of a separating hyperplane at w.

Need for interior

As an example, suppose we believe that the 'market', defined by equal weightings on all assets, will not go either up or down, but we have no opinion about the relative motions of its components. In other words, we believe that $r_1 + \cdots + r_n = 0$. We might attempt to capture this in our framework by setting $D_1 = (1, \ldots, 1)^\mathrm{T}$ and $D_2 = (-1, \ldots, -1)^\mathrm{T}$; the consistent cone is then $Q = \{r = (r_1, \ldots, r_n) | r_1 + \cdots + r_n = 0\}$, which has empty interior. There are *no* pairs of portfolios for which $w \succ v$, and *every* point is efficient. The above theorems do not tell us anything about normals to $\mathcal{M}$.

Appendix B: Computation of centroid vector

Given a wedge domain Q, the centroid c is defined as the geometric centroid of Q, under any radially symmetric density function. Of course, c is defined only up to a positive scalar constant, and hence the radial structure of the density is not important.

Monte Carlo

The simplest way to calculate c is by Monte Carlo. Let x be a sample from an n-dimensional uncorrelated Gaussian, and for the single-sector case, let y be the vector whose components are the components of x sorted into decreasing order. Then $y \in Q$, and since the sorting operation consists only of interchanges of components which are equivalent to planar reflections, the density of y is a radially symmetric Gaussian restricted to Q. The estimate of c is then the sample average of many independent draws of y.

The multiple-sector case is handled simply by sorting only within each sector. Note that this automatically determines the relative weights between the sectors.

The case with comparison to zero is also easily handled. The initial Gaussian vector is sign-corrected so that its first ℓ components are non-negative and its last $n-\ell$ components are non-positive; then a sort is performed within each section. Clearly, each of these operations preserves measure.

For more complicated inequality information structures, the geometry is not always so simple; it is not always possible to reduce a general point x into the wedge Q by measure-preserving reflections. Each new situation must be evaluated on its own.

Direct calculation

For a single sort, computing the centroid is a special case of the general problem of order statistics (David and Nagaraja, 2003). Let x be an n-vector of independent samples from a distribution with density $f(x)$ and cumulative distribution $F(x)$; in our case this density will be a standard Gaussian so that the density of x is spherically symmetric. Let y be the vector consisting of the components of x sorted into decreasing order. Then elementary reasoning shows that the density of the jth component $y_{j,n}$ is

$$\text{Prob}\{w < y_{j,n} < w + \mathrm{d}w\} = \frac{n!}{(j-1)!(n-j)!} F(w)^{n-j}(1-F(w))^{j-1} f(w)\mathrm{d}w$$

The centroid component $c_{j,n}$ is the mean of this distribution:

$$\begin{aligned} c_{j,n} &= \frac{n!}{(j-1)!(n-j)!} \int_{-\infty}^{\infty} w F(w)^{n-j}(1-F(w))^{j-1} f(w)\mathrm{d}w \\ &= \frac{n!}{(j-1)!(n-j)!} \int_0^1 F^{-1}(z) z^{n-j}(1-z)^{j-1}\mathrm{d}z = \mathbb{E}_g(F^{-1}(z)) \end{aligned} \tag{4.8}$$

where $\mathbb{E}_g(\cdot)$ denotes expectation under the probability density

$$g(z) = \frac{n!}{(j-1)!(n-j)!} z^{n-j}(1-z)^{j-1}$$

When j and n are large, this distribution is narrow. Thus, reasonable approximations to the integral are either $F^{-1}(z_{\text{mean}})$ or $F^{-1}(z_{\max})$, where the mean and the peak of the distribution are

$$z_{\text{mean}} = \frac{n-j+1}{n+1}, \; z_{\max} = \frac{n-j}{n-1}$$

(Using the max value has the disadvantage that it requires $F^{-1}(z)$ at $z = 0, 1$, which is not defined.)

For the normal distribution, these formulas are special cases, with $\alpha = 0, 1$, of 'Blom's approximation' (Blom, 1958)

$$c_{j,n} \approx N^{-1}\left(\frac{n+1-j-\alpha}{n-2\alpha+1}\right)$$

Blom shows (in part analytically, in part reasoning from numerical results) that the values $\alpha = 0.33$ and 0.50 provide lower and upper bounds for the true value of $c_{j,n}$, and he suggests that $\alpha = 0.375$ is a reasonable approximation for all values of j, n. More detailed calculations (Harter, 1961) suggest that values closer to $\alpha = 0.40$ give more accurate results when n is large.

By comparison with numerical integration of equation (4.8), we have found that an excellent approximation is $\alpha = A–Bj^{-\beta}$, with $A = 0.4424$, $B = 0.1185$, and $\beta = 0.21$. This gives centroid components with maximum fractional error of less than 0.5% when n is very small, decreasing rapidly as n increases. Since c is defined only up to a scalar factor, the errors in the normalized coefficients will be smaller.

For multiple sectors, the above procedure may simply be applied within each sector. Because we have been careful to preserve the normalization, the relative magnitudes are correct.

Acknowledgements

We are grateful for discussions with, and comments from, Steve Allen, Clifford Asness, Alex Barnard, Michael Chong, Sam Hou, David Koziol, Andrei Pokravsky, Matthew Rhodes-Kropf, Colin Rust, and especially Sebastian Ceria and Robert Stubbs of Axioma Inc.

References

Almgren, R. F. (2003). Optimal execution with nonlinear impact functions and trading-enhanced risk. *Applied Mathematical Finance*, 10:1–18.

Almgren, R. and Chriss, N. (2000). Optimal execution of portfolio transactions. *Journal of Risk*, 3(2):5–39.

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics*, 9:3–18.

Blom, G. (1958). *Statistical Estimates and Transformed Beta Variables*. New York, NY: John Wiley and Sons, pp. 69–72.

Boyd, S., and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge: Cambridge University Press.

Campbell, J., Grossman, S. J. and Wang, J. (1993). Trading volume and serial correlation in stock returns. *Quarterly Journal of Economics*, 108:905–939.

Daniel, K. and Titman, S. (1997). Evidence on the characteristics of cross sectional variation in stock returns. *Journal of Finance*, 52:1–33.

Daniel, K. and Titman, S. (1998). Characteristics or covariances: building superior portfolios using characteristics. *Journal of Portfolio Management*, 24:24–33.

David, H. A. and Nagaraja, H. N. (2003). *Order Statistics*, 3rd edn. New York, NY: John Wiley and Sons.

Fama, E. F. and French, K. R. (1992). The crosssection of expected stock returns. *Journal of Finance*, 47:427–465.

Fama, E. F. and French K. R. (1996). Multifactor explanations of asset pricing anomalies. *Journal of Finance*, 51:55–84.
Harter, H. L. (1961). Expected values of normal order statistics. *Biometrika*, 48:151–165.
Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7:77–91.

5 Some choices in forecast construction

Stephen Wright and Stephen Satchell

(*The views and opinions expressed in this article are those of the authors and are not necessarily those of UBS. UBS accepts no liability over the content of the article. It is published solely for informational purposes and is not to be construed as a solicitation or an offer to buy or sell any securities or related financial instruments.*)

Abstract

Realistic forecasts merge information from many sources. This information may be subjective or objective; absolute or relative; strategic or tactical; incomplete or contradictory; macroeconomic, factor, style or stock specific. No matter how complex, we need a coherent framework to allow the insight from these multiple inputs to be merged in a way that is both statistically rigorous and intuitively reasonable.

In reality, forecasts of return are probability distributions that can become arbitrarily complex (multivariate and/or multimodal) if they are to fully reflect all the above insight. Unfortunately there is no single 'correct' way of constructing such a distribution; the most effective approach depends on the characteristics of the system being forecast, as well as the information quality and skill available to the forecaster.

This chapter builds on our earlier paper 'A robust cross-sectional approach to equity forecast construction' (Satchell and Wright, 2005), which introduced the concept of a mixture of normals approach in the context of rank scorecards. Here we extend this approach to show how insight from equilibrium models, relative value views, expected factor return, and expected asset return as well as stochastic scenario forecasts can all be combined in the same framework. We then show how it can be further extended to reflect the effects of including assets with an asymmetric payoff function (e.g. convertible bonds, options and structured products, etc.).

The problem with this type of approach is that it can result in highly non-normal expected return distributions that are not handled well by classical mean variance analysis. However, we show that a mean-CVaR optimization approach agrees with the classical model when the forecast is a normal distribution, while giving more reliable recommended holdings when the underlying expected return distribution is in reality non-normal.

5.1 Introduction

Forecasting investment returns is not an easy task. To the extent that the underlying economic processes are deterministic, they are usually non-stationary, multivariable

and highly non-linear. They are also plagued with measurement problems such as the inconsistent reporting standards for company accounting data in different jurisdictions, or the frequent revision of macroeconomic data series by national statistics offices. Even when the data are accurate and consistent, the reporting may be infrequent and delayed or the available time series may be very short.

Our view on financial markets is that asset returns are highly state-dependent non-linear functions of underlying exogenous variables wherein the key parameters, if they can be identified, exhibit profound degrees of time-dependence. Faced with such a bleak description, it may be thought that quantitative finance has little to offer; however, these problems do not affect all investors equally, and this is what drives the diversity of forecasting and risk management strategies that can be observed in the marketplace.

For example, it is of interest to some long-term 'buy and hold' investors to establish equilibrium states toward which the economy can be expected to mean revert. For these purposes we can use large sample theory to mitigate some of the above difficulties, because it allows us to make simplifying assumptions such as stationarity and ergodicity.[1] These properties allow us to get closer to the true, unknown data-generating process simply by observing long enough time series, because the departures from this simplified model become 'lost in the noise'.

Another investment style is driven by the quarterly, semi-annual and/or annual assessment of a fund manager's performance. For this group the transient behaviour of the market and how it moves between states is all-important, as this is what generates short-term investment returns. For these investors, 'one-off' events and non-linearities are of dominant importance because they are what disturb the equilibrium (if it is ever fully achieved following earlier exogenous shocks to the system). Hence for this class of investors we need to move from a frequentist view of modelling to a likelihood view where we can augment the available noisy data with logical extrapolation, using theory, judgement and experience.

In addition to these differences in the time horizon between investors, there exist many other differences of investment approach depending on the particular skill of each individual or organization. Some investors are specialists in top-down or macroeconomic analysis. Their role is to pick asset classes, regions or sectors that are likely to outperform, based on the large-scale economic factors. Others believe that their skill lies in bottom-up stock picking, based on careful analysis of each company's market, management and product strategy.

Hence we have short- and long-term, top-down and bottom-up investors. To this we can add relative or absolute, and quantitative versus technical approaches to investment management. The reality is that all these dimensions are important in any forecasting process, and either they should be accommodated in it, or the risks implicit in not forecasting a dimension should be hedged away.

To further complicate this picture, the importance of each style of forecasting varies over time. For example, top-down macroeconomic views are critical at turning points in the business cycle. Stock-specific issues dominate mid-cycle, when the macroeconomic environment is benign and predictable.

[1] Intuitively, a stochastic process is ergodic if its sample properties mirror its population properties. For example, we might expect the average sample return to be close to the unknown population return.

An idealized investment strategy therefore needs to be tailored to the detailed characteristics of each investor, and then varied continuously as circumstances change. Of course this is much easier to suggest than it is to do in practice, because of the ever-changing kaleidoscope of factors and one-off events that drive investment returns. However, what we can do is provide a framework within which investors can exploit a mixture of objective analysis for those aspects that are amenable to quantitative modelling combined with the theory, experience and judgement, cited above, for those aspects that are not. Our objectives are to make the most of all available information, record and discuss our views as concisely and realistically as possible, avoid inconsistencies and inefficiencies in our implementation, manage our risk exposure as realistically as we can, and learn from our experience to maximize our performance in the future.

We do not claim to be able fully to meet all these aspirations; they do, however, provide the context for this chapter, which is primarily concerned with one stage in this process design – that of forecasting. Our contention is that investment models can be designed to have certain properties, and that when you approach the task from this perspective you arrive at some very interesting conclusions that are of crucial importance to the practising investment professional.

This chapter follows on from previous work where we used rank information in portfolio construction (see Satchell and Wright, 2003). There, we showed that this dealt with many of the difficulties encountered when building portfolios using knowledge of forecast alphas.

In a later paper, we extended this rank approach to include forecast construction where the resolution of the aggregation both in return space and across your classification structures is related to the quality of the information contained in that forecast (Satchell and Wright, 2005). In particular, in that paper we showed how this rank scorecard approach can be related to a classic linear factor structure model and its generalization into a mixture of normals model, hence allowing the rank scorecard to be used as a reduced form model or reporting format for a wide range of more complex models.

In this chapter, we outline some of the more common model forms used by practitioners and relate these to this mixture of normals approach. The problem with this type of approach is that it can result in highly non-normal expected return distributions that are not handled well by classical mean-variance analysis. However, we show that a Mean-Conditional-Value-at-Risk (CVaR) optimization approach agrees with the classical model when the forecast is a normal distribution, while giving more reliable recommended holdings when the underlying expected return distribution is in reality non-normal.

In Section 5.2, we review the fundamental building blocks for our approach – the linear factor model; this is central to all quantitative portfolio analysis. Another important notion is what we could call 'local normality'; this is not to claim that returns are normal, but that returns over short periods of time and/or conditional upon particular states of the world are normal. Such an approach allows for unconditional returns being non-normal; it also encompasses log normality and the style of analysis used in the option-pricing literature.

In Section 5.3 we consider issues of modelling the returns distribution as a mixture of normals; in Section 5.4 we cover practical issues in constructing models as a mixture of normals. Section 5.5 discusses the mean-CVaR approach to optimization given a Monte Carlo distribution, and we present our conclusions in Section 5.6.

5.2 Linear factor models

In this section, we first review linear factor models. Let y_{it} be the return/rate of return to asset i at time t. Let β_{ijt} be the exposure of the return of asset i to factor j at time t. Let f_{jt} be the value of factor j at time t. We assume n assets, $i = 1 \ldots n$. We assume k factors, $j = 1 \ldots k$. The linear factor model (LFM) takes the following form:

$$Y_{it} = \sum \beta_{ijt} f_{jt} + V_{it}, \qquad i = 1, \ldots, n, j = 1, \ldots, k \tag{5.1}$$

The error terms V_{it} represent the idiosyncratic component of the LFM. This is usually modelled as a random variable, V_{it}, which has a mean α_{it} and a variance σ_i^2. The key assumption of equation (5.1), other than linearity, is that it is the factors f_{jt} that they pick up the common covariation in y_{it}, so that the V_{it} are uncorrelated. By use of vectors and matrix notion, equation (5.1) can be rewritten as

$$y = \beta f + V \tag{5.2}$$

where y and V are $(n \times 1)$ vectors, f is a $(k \times 1)$ vector, β is an $(n \times k)$ matrix.

It is clearly unlikely that such an attractive structure is an accurate description of reality, however much we may wish it to be so. Our perspective on this is that of the mathematician interested in linearizing highly non-linear dynamic equation systems. We are motivated by the fact that, however complex and non-linear the real world may be, the operational decision facing the investment committee is always the same – i.e. 'what do we do next, given our understanding of current local linearities?' A detailed discussion of such procedures can be found in Dutton *et al.* (1997); here, we shall content ourselves with an overview.

Generally, suppose that, following the notation of equation (5.2)

$$y = H(f, U) \tag{5.3}$$

Here H is a $(n \times 1)$ vector non-linear function of f and U, where U is some $(m \times 1)$ source of noise in the system.

By taking the usual Taylor's-series expansions around zero vectors, we see that

$$y = H(0, 0) + H1(0, 0) f + H2(0, 0) U + \text{Rem}t \tag{5.4}$$

where $H1$ (0, 0) is the $(n \times k)$ matrix of partial-derivatives with respect to f, $H2$ (0, 0), is the $(n \times m)$ matrix of partial derivative with respect to U; both are evaluated at zero, and Rem$t$ corresponds to higher-order terms in the expansion. To see the links between equations (5.4) and (5.2), we equate β_t to $H1$ (0, 0) and similarly,

$$V = H(0, 0) = H2(0, 0)\, U + \text{Rem}t \tag{5.5}$$

It is immediate from the above that we can incorporate more complex schemes than equation (5.3), i.e. where y is implicitly connected to f.

Furthermore, we can (and probably would want) to expand the Taylor's series around some point rather than (0, 0).

It is worth enquiring as to the origins of equation (5.3); where should such a complex structure come from? The short answer is that if we start with data-generating processes that are, in the short run, exogenous to the market, such as productivity, climate, etc., and we know the utility/decisions functions of all individuals and organizations that participate in the model, we can compute an equilibrium which will link return distribution parameters to utility characteristics. These will take the form of highly non-linear restrictions which can be substituted back into the data-generating processes to arrive at an equilibrium pricing process, which is equation (5.3).

Equation (5.3) will also encapsulate all the heterogeneous assessments of probability and behavioural quirks of the participants, not to mention the current state of institutions. An example of the above which is still highly simplified is provided by Cox *et al.* (1985).

Inspecting the previous argument, we see that local linearity seems both reasonable and plausible except that:

1. The point about which the linearization occurs will change with time
2. The partial derivatives that become the coefficients in the linear system will change their values as the points in (1) change.

Using linear models successfully requires awareness (1) and (2) above. In particular, procedures that assume that coefficients in the linear factor model are fixed are likely to suffer from severe inaccuracies through time, unless the underlying structure is actually linear. Practitioners are aware of these difficulties and build their models using rolling windows; this means that models will change their linear coefficients through time. However, this will not always work.

Virtually all linear factor models failed in the late 1990s due to their inability to pick up the Internet factor. With a rolling window of, say, 7 years of monthly data, it will take several years for the Internet factor to manifest itself in the model. Even when it does, the factor may have then ceased to be of any importance. The industry response to this has been to move to higher frequency data to make the models more responsive to evolutionary changes. This brings its own problems of increased sensitivity to noise and failure to accommodate the fact that these parameters may not vary smoothly over time, but exhibit jumps from one value to another.

Furthermore, measuring at high frequency measures something different, i.e. correlation of hourly data is not the same as correlation of monthly data because, for example, the hourly data will pick up the end-of-day effects where traders need to leave their books balanced overnight. These dynamics are entirely absent from monthly data, and hence we can see the technical issue, which is that of temporal aggregation; it is not always possible to recover the density of daily prices by convoluting the density of hourly prices.

The approach that we advocate essentially allows the factor exposures to be recalculated every period while being agnostic as to how the recalculation is done. This accommodates all three types of assumption at the user's discretion, i.e.

1. Stationary parameters
2. Smoothly varying parameters over time
3. Discontinuous parameter behaviour.

In this way our model structure can be thought of as a way of summarizing the conclusions about 'the state of the market' at each moment in time, derived from a much

wider set of possible model forms either explicitly quantitative in form or implicitly so, having being constructed by a mixture of theoretical inference and quantitative support. In fact, it can be shown how a range of modelling techniques that are often thought of as independent of each other actually lie on a spectrum of increasing complexity as your simplifying assumptions are relaxed from (1) to (3) above.

Some readers may at this point conclude that we are in some sense cheating in that our proposed model form throws no additional light on how the markets work. However, our contention is that the only certainty in forecasting is change, and the only decision that is necessary at each re-balancing point is what to do next. From our perspective, we wish to design an investment process that accommodates the former and illuminates the latter in a simple understandable manner. Hence, we are agnostic as to which detailed model structure may best model market conditions over time, or even which model is currently performing best.

What we aim to provide is a standard form that summarizes (and allows the user to compose and/or merge) the conclusions from different models in a systematic and statistically rigorous manner while still being widely intelligible to all the participants in the investment decision – i.e. our approach is a framework for comparing, contrasting and communicating the results of other models, rather than a model in its own right.

5.3 Approximating risk with a mixture of normals

In this section we discuss return distributions. The usual assumption of multivariate normality for conditional or unconditional returns does not strike us as reasonable, although it may hold at the level of well-diversified portfolios with a long investment horizon due to the operation of the central limit theorem.

We prefer to assume that conditional expected returns are multivariate normal, conditionally for each single period. We could weaken this assumption to allow returns to follow mixtures of multivariate normals. Thus we can structure the problem by assuming that there are S states of the world, and conditional upon each state we get a multivariate normal distribution but with different means and variances. Practitioners are probably aware of this framework without the necessary formal mathematics. A simple but important example might be where $S = 4$, the four states being:

1. High alpha, high risk
2. High alpha, low risk
3. Low alpha, high risk
4. Low alpha, low risk.

Each such state would have a probability attached to it. It may be possible to identify the states, as we could in the above example, or the states may be hidden, as in the examples of hidden Markov models or, as they are more frequently known, regime-switching models (see Hamilton, 1989). Arguably, in the above example we are currently in a low-alpha, low-risk environment (regime 4). However, *circa* 2000–2002 we were in a low-alpha, high-risk environment (regime 3).

Formally, our returns model could be described by the following:

Suppose there are m regimes, $l = 1 \ldots m$. For each regime, returns are described by the following probability density function (pdf):

$$Pdfi(x) \sim N(\mu_l, \Omega_l) \tag{5.6}$$

where μ_l, and Ω_l are the mean vector and covariance matrix of returns in regime l and $N(.)$ denotes multivariate normality.

For each regime l, there is a regime probability π_l such that

$$\sum \pi_l = 1. \quad \mathrm{l} = 1, \ldots, \mathrm{m}$$

The unconditional returns *pdf* follows a mixture of normals; symbolically,

$$pdf(x) = \sum \pi_l pdf_l(x) \tag{5.7}$$

Such a specification follows naturally in a Bayesian framework with multiple priors, in which case the prior and the posterior follow equation (5.7), assuming that the data are multivariate normal. Another important property of mixture of normal and log-normal distributions is their ability to approximate, with very close accuracy, arbitrary return distributions.

This approximation property is widely used in the options-pricing literature to construct estimates of risk-neutral distributions, and also in the stochastic volatility literature, where many of the joint densities whose formulation is not known in closed form are replaced by low-order mixtures of normals or log-normals. From the perspective of a cross-sectional linear factor model regression, all of the above add no complexity to the econometrics as long as we are implicitly or explicitly conditioning on the regime information.

Of course, if there are dynamics in the system, this regime information will itself be serially correlated as with GARCH or Kalman filter models (see Diderrich, 1985). Conceptually this is little different to the regime-switching models above, except that the parameters are assumed to move smoothly through a series of states (or adjacent regimes), thus needing a larger number of such states to adequately represent a mixture of normals approximation to the risk. (In practice, of course, this does complicate the parameter estimation as the quantity of data for each state is reduced, but that is a practical consideration outside the scope of this chapter.)

Last but by no means least, we started the discussion in this chapter by talking about the importance of one-off events and the difficulties that these presented to frequentist modellers. In the discussion so far we have, in the main, discussed a model evolving through time, hence generating a mixture of normals. An alternative view is where we may have a number of candidate models but not be confident which applies at any moment in time. If we give each alternative a degree of belief, then we are again in a mixture of normals environment. This also provides an entry into the world of Bayesian statistics where we update our model not just over time, but also from a parsimonious prior to our posterior probability in a cumulative manner as we successively take into account each piece of information (subjective or objective) we hold together with our confidence in that item of information. In this case, we would arrive at equation (5.7).

From this, we can see that the basic mixture of normals formulation provides a standard mathematical form that is compatible with a range of underlying risk models. In particular, it allows us systematically to include subjective information in our analysis and merge it with all our information from other sources. In practical terms, we might have a precise scorecard based on our subjective information which is updated by detailed model information where it is present, and in proportion to the confidence in that model.

Hence, we have achieved for our risk input a similar generalized form which can capture the conclusions on risk from a range of underlying models while being agnostic about each particular source. Unfortunately, the standard form in its most general incarnation is not designed for manual interpretation in that multivariate normal distributions are intimidating when presented in isolation; when presented as a mixture, simply interpreting the numbers to identify the detailed behaviour implied therein becomes unrealistic. Fortunately, because the underlying structure is conceptually simple, flexible and easy to describe, computer-aided tools can come to our assistance by letting the user work with the key investment interpretations while hiding the necessary but confusing detail. This is, after all, routine algebra that can be safely taken as read.

The main advantage of this unifying form is that it allows us to see the strengths and weaknesses of different approaches not as unconnected alternatives but as part of a spectrum with properties that vary smoothly as we move along it, and where we can change our position on this spectrum as our circumstances change over time without needing to radically change our underlying philosophy or support tools.

In practice, we will often adopt some simplifying assumptions to radically simplify this presentation. For example, we might choose to model the four-state model above by having four scorecards, one for each scenario, give each a probability and assume that the covariance structure is the same for each. We can then add increasing degrees of sophistication when (and only when) they are needed to adequately reflect important detail in the underlying problem.

While this mixture of normals risk modelling framework provides an elegant, unified approach to implementation, it cannot absolve the user from the responsibility of thinking about some fundamental practical problems inherent in any model-building process. In the next section we discuss these traps for the unwary, prefaced where necessary by an introduction to the relevant theory.

5.4 Practical problems in the model-building process

5.4.1 Independence of inputs to your forecast

One of the most common ways of model-building is to use simple linear regression to identify some variable that is believed to have been a lead indicator for some asset returns. When we do this, we have the problem that our factors are not always independent of each other. In these circumstances, if we do a stepwise regression the fitted value of the coefficients will depend on the order in which the different terms are introduced. This is because if two factors are highly correlated, whichever is introduced first will pick up most of the variation that is associated with the pair. Hence, the coefficients in your model are not a single unique set of values even when you have comprehensive and accurate data on the system.

When constructing a forecast, you are often explicitly or implicitly using sub-models that have been calibrated independently. For example, your top-down forecasts may be prepared by an economist or strategist, while your bottom-up forecasts are prepared by a set of company analysts or fund managers. Hence the drivers for these sub-models may be highly correlated without you realizing it. If you now enter high scores for two correlated factors, you are in effect double-counting that score.

There are several palliative actions which can be adopted to minimize these problems. The simplest is careful selection of your factor, set to be as independent as possible. More sophisticated approaches range from jointly estimating all the driver sensitivities so that the stepwise regression effect counteracts the double-counting effect, through to the formulation of Grinold and Kahn (1999: 263), where you assume that you have a knowledge of the full covariance between forecast and returns as follows:

$$E(y|g) = E(y) + Cov(y, g).\, Var^{-1}(g)(g - E(g)) \tag{5.8}$$

where g is the vector of raw forecast scores and y is the return vector. In words, this is saying that our expected return given our scores is consensus expected return plus a scaled term proportional to the deviation of our score from its consensus value – the scaling term being derived from the covariance of observed return with the forecast scores.

5.4.2 Independence of updates in a Bayesian context

One of the more elegant forecasting techniques that you can use is based on a Bayesian analysis, where the investor has prior information represented by a probability distribution for likely return. This is then updated as new information arrives to produce an updated distribution.

If the information in the prior and the update are independent of each other, then this results in a very attractive cumulative process which is easy to implement – as we can see by considering Bayes' theorem in more detail. At one level, Bayes' equation is little more than an axiomatic statement about conditional probabilities:

$$P(A|B) = P(A)(P(B|A)/P(B))$$

For our purposes here, this is really an updating equation. Given an initial probability P(A), we can refine that probability, as new information becomes available in the form of an observation with probability P(B), and the likelihood of seeing that data given your hypothesis (P(B|A)/P(B)).

This equation is easily extended to refer to probability distributions rather than probabilities. For this to be computationally convenient, we need to assume a form of these distributions. Within financial forecasting circles, the assumed distribution form will almost always be multivariate normal.

When we multiply a multivariate normal distribution (representing our prior views) by another (representing our likelihood), the result (our posterior probability distribution) is itself a multivariate normal. Hence it can be used in its turn as a new prior to which you add further information. This is called a conjugate prior form.

In principle, we can carry on in this manner indefinitely; however, there is a built-in assumption in so doing that we must recognize, and that is that at each stage the new

information distribution is independent of all of the previous inputs. An example might clarify this idea of independence.

Suppose you have entered a forecast return distribution for a company that was based at least in part on the assumption about its sector going into recession with poor sales prospects. Now you want to update this by adding some company-specific information; they have just appointed a well-respected MD and you believe this will affect their prospects relative to the sector. Is this independent information or not?

If the new appointee is well respected because he is an excellent salesman rather than good at controlling costs, this new information is NOT independent. The extent to which you expect his skills to improve the company's performance depends on whether the company finds itself in an environment where sales growth or cost-cutting are the key skill areas. When new information is not independent, you can still add it. However, this needs to be done in such a way that the related items of information are simultaneously input in the form of a joint probability distribution. This allows you to include the correlation between the management change and the expected macroenvironment.

In principle, you could assume that all your input was correlated with all of the other inputs. If so, you would then need to estimate all of those correlations. If you follow this route, you run the risk of introducing new potential sources of noise in your calculation. Doing this can make it increasingly difficult for you to maintain a physical insight into the estimation process. This will have been replaced with a dependence on the accuracy and stability of these intermediate correlation calculations. It can now be seen that this is essentially the same problem as identified above for the Grinold and Kahn equation.

Of course, these issues are never black and white. If the assumption of independence is a sufficiently good approximation, then it allows you to keep a simpler, more understandable model. If it is not a good approximation, then there should be a strong relationship that can be reliably captured in a correlation coefficient. In practice, this is the heart of the model-building process. Models are approximations of reality. Model-building is the process of deciding when these approximations are sufficiently close, and when you need to elaborate the model.

5.4.3 Relative value, factor and scenario information

Frequently when constructing a forecast, the insight that is available will not be an expectation on absolute return, but one of likely performance of one asset relative to another with an estimated uncertainty around that value. At any point in time there may be many such views on likely relative performance that we wish to include in our analysis.

The Bayesian updating framework that we have considered in the last section can be conveniently extended to cater for this problem. This is then known as the mixed estimation approach to establishing a relative value forecast, as popularized by Black and Litterman (BL). This takes a prior multivariate normal distribution (in the BL formulation this prior is derived from a set of CAPM equilibrium assumptions) with a mean of μ and covariance S, i.e.

$$pdf_1(r) = c\exp(-0.5(x-\mu)'S^{-1}(x-\mu)) \tag{5.9}$$

It then updates it with a second multivariate normal

$$pdf_2(x) = c\exp(-0.5(x-g)'V^{-1}(x-g)) \tag{5.10}$$

where x is the relative return (constructed by calculating the return on an arbitrary number of linear combinations of return on each asset) with a mean of g and covariance of V (this can be thought of as the mean and covariance of return on a set of hypothetical portfolios). That is,

$$x = P.r \tag{5.11}$$

Substituting (5.11) into (5.10) gives:

$$pdf_2(r) \ \alpha \ exp(-0.5(P.r - g)'V^{-1}(P.r - g)) \tag{5.12}$$

Remembering that $e^a \times e^b = e^{a+b}$, the product of the two multivariate normals equations (5.1) and (5.4) can be written as:

$$pdf_1(r) \times pdf_2(r) \ \alpha \ exp(-0.5(r - \mu)'\mathrm{S}^{-1}(r - \mu) - 0.5(P.r - g)'V^{-1}(P.r - \mathrm{g})) \tag{5.13}$$

In order to return this to standard multivariate normal form (and hence achieve conjugate prior form), we need to expand the quadratic elements of this equation and collect terms. This results in the following equations for the updated mean β and covariance Σ:

$$\beta = [S^{-1} + P'V^{-1}\ P]^{-1}[S^{-1}\mu + P'V^{-1}\ g]$$

$$\Sigma = [S^{-1} + P'V^{-1}P]^{-1}$$

(The details of this expansion and collection of terms is given in Satchell and Scowcroft (2000) and reproduced in this book.)

While this process uses a Bayesian idea of taking a prior and updating it with a likelihood function, it is important to recognize that this likelihood specified by the model builder is the likelihood of these relative values being true conditional on your other relative views also being satisfied. In the general case, this condition is likely to be satisfied at the mean of the distribution but not elsewhere. Hence, the covariance matrix resulting from this analysis should not be interpreted as representing the real expected covariance of the likely asset return. Also, for the estimation of the mean, the result can sometimes be unstable to small changes in the input parameters.

These effects are best understood by taking a stylized example of perfect confidence in our relative value terms. If we do this, we can see that the effect of the prior distribution becomes negligible, and the updating terms become in effect a set of simultaneous linear equations. Hence we should expect that under some circumstances we will suffer from ill-conditioning in this analysis.

The strengths of this approach are its practical operational benefits. Taking the above comparison with simultaneous linear equations, we can now specify a degree of confidence in each equation and find the best fit, both given the different confidence value, and also biased by the prior toward the most likely solution.

This has a number of practical consequences:

1. We will be able to find a solution and a confidence in that solution
2. Inconsistent inputs will be reconciled

- if they are equal confidence, their effects will be the average of the two independently
- if one is high confidence, it will dominate the low confidence input

3. If a forecast is missing from the update set of 'fuzzy' simultaneous equations, then this value will be filled in from the prior.

Hence this is a very convenient way of entering multiple relative value views; however, the same basic analysis can also be used to enter factor views. The key to this is to observe that in equation (5.12) above, x can be allowed to represent the asset return (rather than the relative return) and r to represent the factor return (rather than the asset return). Because of the conjugate prior structure of these equations, we can therefore merge factor views with our asset and relative value views by repeatedly using the same basic equation.

The final type of insight that we wish to incorporate into our mixture of normals approach is views on scenarios and scenario probability. In these models, a series of discrete states are postulated to represent the expected distribution of return. As with the other forecasting techniques, the independence or otherwise of these scenarios needs to be carefully considered and a sufficient number generated to allow any optimization process to identify a realistic risk-return tradeoff. Typically, this requires many more scenarios than there are assets.

As the number of such individual scenarios increases, the task of generating them becomes more tedious. Hence our formulation of stochastic scenarios, where we assume states of the world each with its own probability distribution in multivariate normal form from which a set of individual scenarios can be drawn – i.e. in our formulation, scenarios are the sum of a number of normal distributions with different probability.

In this way we have arrived at the general form of our mixture of normals approach where the resulting non-normal distribution can be represented by a Monte Carlo distribution. When we do this, it is then trivial to see that we can apply arbitrary payoff functions to this distribution to capture the effects of optionality, convertible bonds or structured products. The problem then is to optimize, given this Monte Carlo set, as discussed in the next section.

5.5 Optimization with non-normal return expectations

The problem with a mixture of normals approach is that it can result in highly non-normal expected return distributions that are not handled well by classical mean-variance analysis for the simple reason that the Classical MV approach will implicitly fit the nearest normal distribution to the expected return, hence losing all the information contained in the non-normal component of the distribution.

If the departure from normal distribution is important, then alternative metrics need to be used to report risk in the portfolio. For small amounts of non-normality, robust measures like mean absolute deviation or rank correlation might be considered. However, these are both symmetric measures of risk. For highly asymmetric expected

return distributions, you need to consider a threshold measure of risk. The three best known are:

1. Value at risk (VaR), which relates the probability of falling below a return threshold to the level of that return threshold
2. Mean excess loss (also known as conditional value at risk, CVaR), which multiplies that downside probability by the average amount by which the threshold is exceeded
3. Loss gain ratio, which takes that weighted downside value and divides it by the weighted upside.

Each of these approaches has its own strengths and weaknesses. The best known is value at risk (VaR). This identifies the confidence that return on a portfolio over a selected period will be above a given threshold.

There are three major problems with VaR as a risk measure:

1. It does not take into account the distribution conditional on the threshold having been exceeded, therefore leaving the investor uninformed about the severity of any possible long tail risk
2. The VaR for sub sets of the portfolio do not sum to the VaR for the portfolio, which makes diagnosing which are the most risky positions within the portfolio less intuitive
3. Optimization using VaR is a mixed integer optimization problem, which is very computationally expensive to calculate.

The next well-known threshold risk measure is CVaR (or mean excess loss). This measure describes the average expected loss conditional on a return threshold being exceeded. This measure is sub-additative and convex (see Artzner *et al.*, 1999). In particular it is always greater than VaR, hence reducing the maximum CVaR within a set of portfolios will also reduce the maximum VaR. In practical examples, optimizing on CVaR has been found to provide very close to optimum VaR solutions while being far more easily calculated because it can be optimized using a linear programming algorithm, as we shall see below (see also Uryasev, 2000), hence it addresses all the major weaknesses of VaR listed above.

However, from an investment point of view CVaR misses an important component of the problem in that given two portfolios with the same CVaR, a rational investor would prefer whichever portfolio had the greater upside potential – hence the final one of our threshold measures, the loss gain ratio. This divides CVaR by the upside potential measured as the weighted probability of exceeding the threshold return. The disadvantage of this final measure is that it reflects the optimality of your portfolio in risk per unit return, rather than providing a pure (downside) risk measure. In practice, most investors would want to see both the CVaR measure and the loss gain ratio. Because VaR is such a widely quoted risk measure, this may well be the statistic to report even though the other two are the ones used for portfolio selection.

Luckily, all three threshold risk measures are easily computed as part of the CVaR optimization process. Here we are assuming that we are optimizing based on a set of Monte Carlo simulated time series.

At each point in time, each asset has a simulated return – hence, if we consider the weight space encompassing all our possible portfolios, we can draw a hyperplane in this asset weight space where all portfolios exceeding the threshold return lie above this hyperplane while all others lie on the other side.

The simplex linear programming method solves a problem of the form:
Maximize the function

$$Z = a_1x_1 + a_2x_2 + \cdots a_nx_n$$

subject to

$$x_i \geq 0, i = 1, \ldots n.$$

and i constraints of the form

$$a_{i1}x_1 + a_{i2}x_2 + \cdots a_{in}x_n \leq b_i$$

or

$$a_{i1}x_1 + a_{i2}x_2 + \cdots a_{in}x_n \geq b_i$$

or

$$a_{i1}x_1 + \cdots a_{i2}x_2 + a_{in}x_n = b_i$$

These constraints also form hyperplanes in weight space. These effectively bound a feasible region within which solutions are allowed. This region can be shown to be convex, so the maximum feasible value of the objective function occurs at a vertex on its boundary. The algorithm proceeds by finding an initial feasible solution and then finding a better adjacent vertex, moving to this and then repeating until no further improvement is possible.

Unfortunately our initial problem formulation is similar but not identical to this LP standard form. However, some simple tricks can convert one to the other. The main difficulty is that we can have viable solutions on either side of our hyperplanes in weight space. If we consider one such hyperplane, we can add two dummy variables (say d^+ and d^-) to our problem definition and constrain the values of these dummy variables such that they represent positive and negative distances above and below this hyperplane respectively. These are now firm constraints on the dummy variables, so bounding the augmented feasible region. Replacing all our original constraints with new ones in this form allows the objective function to be the sum of distances below each hyperplane. Hence minimizing this minimizes CVaR.

In this formulation it is also easy to count the number of hyperplanes above and below any point in weight space and to evaluate the upside risk as well as the downside risk. Hence calculation of the VaR and loss gain ratio is a simple extension of this approach, even if neither is easy to optimize directly; VaR because it is an integer function (counting the number of hyperplanes above any given point in weight space), and the loss gain ratio because it is a ratio of the upside and downside risk, so CVaR as well as being the easiest to interpret risk measure is also the only linear measure that we can optimize using this approach.

5.6 Conclusion

In our view, equity returns are driven by an ever-changing mixture of factor influences, one-off events and idiosyncratic (asset-specific) issues. Put more strongly, we believe that there never can be one unique 'correct' solution to this forecasting problem, because as the nature of current return drivers becomes more understood they are then increasingly arbitraged away. Hence our modelling process is inevitably chasing a moving target.

Secondly, even at one point in time, different investors, with different skill sets and circumstances, will rationally choose to forecast very different aspects of the market (e.g. relative rather than absolute return, or top-down rather than bottom-up return) while simultaneously hedging out those other aspects which they are not seeking to forecast.

This has caused us to focus on how to design an investment process or modelling framework to organize your thoughts on these issues rather than seeking one specific model. From this perspective, the key issues are the flexibility to reflect the different model forms, robustness to noise in the data, simplicity, parsimony and an intuitive reporting format to aid comprehension and communication of the results, combined with theoretical rigour to maximize confidence in the output of the process.

Our earlier paper (Satchell and Wright, 2005) proposed the use of order statistics as a key input to your forecasting process and related this to mainstream linear factor theory so that the full power of this prior art can be exploited. In particular, it introduced the concept of a mixture of normals approach to risk modelling and the benefits of being able to combine subjective and objective information in a rigorous framework.

This chapter extends this work by showing that the same mixture of normals approach can also support (and merge the results from) a range of alternative model types that have proved popular with practitioners. These are:

1. Scorecards
2. Equilibrium models
3. Factor models
4. Scenario (and stochastic scenario) models
5. Bayesian updating models
6. Relative value models.

By providing a framework within which investors can exploit a mixture of model forms, we can use objective analysis for those aspects that are amenable to quantitative modelling, combined with theory, experience and judgement for those aspects that are not. We may never eliminate uncertainty in the forecasting process, but what we can do is to create tools that allow us to manage that uncertainty as effectively as possible.

References

Artzner, P., Delbaen, F., Eber, J. M. and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9:203–228.

Cox, J. C., Ingersoll, J. E. and Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica*, 53:385–407.

Diderrich, G. T. (1985). The Kalman filter from the perspective of Goldberger–Theil estimators. *American Statistician*, 39(3):193–198.

Dutton, K., Thompson, S. and Barraclough, B. (1997). *The Art of Control Engineering*. Englewood Cliffs, NJ: Prentice Hall.

Grinold, R. C. and Kahn, R. N. (1999). *Active Portfolio Management*, 2nd edn. New York, NY: McGraw-Hill.

Hamilton, J. D. (1989). Analysis of time series subject to changes in regimes. *Journal of Econometrics*, 45:39–70.

Satchell, S. and Scowcroft, A. (2000). A demystification of the Black–Litterman model: managing quantitative and traditional construction. *Journal of Asset Management*, 1:148.

Satchell, S. and Wright, S. (2005). Robust cross-sectional factor modelling approach to equity forecast construction. *Economic and Financial Modelling*, 12(4):153–182.

Satchell, S., Wright, S. and Hwang, S. (2003). Assessing the merits of rank-based optimisation for portfolio construction. In: S. Satchell and A. Scowcroft (eds), *New Advances in Portfolio Construction and Implementation*. London: Butterworth-Heinemann.

Uryasev, S. (2000). Conditional value-at-risk: optimization algorithms and applications. *Financial Engineering News*, 14:1–5.

6 Bayesian analysis of the Black–Scholes option price

Theo Darsinos and Stephen Satchell

Abstract

This chapter deals with the distributional properties of options prices. In particular, we investigate the statistical properties of the Black–Scholes option price within a Bayesian framework. We extend Karolyi's (1993) work to deriving the prior and posterior densities of a European call option by incorporating randomness not only in the volatility of returns but also in the underlying asset price. Numerical results are presented to compare how the dispersion and shape of the option price distribution changes in the transition from prior to posterior information, where information may be price or sample variance, or both. We find that the asset price is very informative in determining the posterior density of the call price. The derived analytical expression for the posterior density is of considerable interest since it can be straightforwardly combined with a loss function to produce optimal estimates of options prices or provide a direct platform for quantile and value-at-risk calculations.

6.1 Introduction

The focal point of a great deal of econometric work within the framework of option valuation has long been the problem of estimating the parameters of continuous-time price processes which act as inputs for parametric derivative pricing models. Since the volatility of the asset price is *conditionally* the only unobservable and potentially stochastic component entering the Black–Scholes (1973) formula, attempts of academics and practitioners to improve on their estimates of option prices have generally focused on the issue of volatility modelling. From the simple models of estimation from historical price and return data (e.g. maximum likelihood using continuously compounded returns calculated from closing prices, Parkinson's (1980) 'extreme value method' incorporating high–low prices, Garman and Klass (1980) adding opening prices, etc.) to ARIMA modelling of the time-series behaviour of volatility (Poterba and Summers, 1987; French *et al.*, 1987),[1] and from the popular (G)ARCH class of stationary conditionally-heteroscedastic processes implicitly allowing for the conditional variance to be time-varying (Engle, 1982; Bollerslev, 1986; Engle and Bollerslev, 1987; Nelson, 1991; Day and Lewis, 1992; Engle and Mustafa, 1992; and many others) to continuous time stochastic variance models (Hull and White, 1987; Wiggins, 1987; Melino and Turnbull, 1990; Scott, 1991; etc.), to implied volatility approaches (Latane and Rendleman, 1976; Chiras and Manaster, 1978; Day and Lewis, 1988). The plethora of practices is overwhelming indeed.

[1]These models act as approximations to a slowly changing time-varying volatility.

It is not the aim of this chapter to antagonize this vast literature with yet another volatility predictor. Rather our scope is to derive, by means of a Bayesian analysis, the 'true' distribution of the Black–Scholes option price (BS hereafter) and thus provide a platform for flexible[2] option price prediction and other risk management calculations such as value-at-risk. The BS price as an unconditional random variable 'actively' depends on both the volatility of returns and the stock price process while as a conditional variable depends only on volatility. One should therefore combine a prior density for the (price, volatility) vector together with the likelihood of volatility to obtain the posterior density of price and volatility. The posterior density of the option price then follows after dividing by the conditional (on the sample and prior information) density of the asset price and applying a non-linear transformation.

Bayesian methods have been used in the past to model the variance of stock returns for the purpose of option valuation. Karolyi (1993) utilizes[3] prior information extracted from the cross-sectional patterns in the return volatilities for groups of stocks sorted by size, or by financial leverage, or by trading volume, together with the sample information, to derive the posterior density of the variance. He reports improved prediction accuracy for estimates of option prices calculated using the Bayesian volatility estimates relative to those computed using implied volatility, standard historical volatility, or even the actual *ex post* volatility that occurred during each option's life. We extend his analysis by deriving the posterior probability distribution of the option price, not just as a transformation of the posterior density of the variance but by incorporating randomness in the underlying asset price as well. Lately there appears to be increasing evidence that incorporating randomness in the underlying asset price in conjunction with randomness in volatility is important and can improve upon theoretical and *ad hoc* models of option pricing (see, for example, Longstaff, 1995; Garcia *et al.*, 2001).

More recently, Bauwens and Lubrano (2000) show how option prices can be evaluated from a Bayesian viewpoint using a GARCH model for the dynamics of the volatility of the underlying asset. Their methodology delivers (via a numerical algorithm) the predictive distribution of the payoff function of the underlying. Our chapter differs from Bauwens and Lubrano's paper in that we follow the log-normal Black–Scholes structure (which allows us to derive the posterior distribution of the option price in analytic form), whilst they follow a GARCH discrete-time structure similar to the one suggested in Duan (1995).

Ncube and Satchell (1997) investigate the properties of the BS price under the classical approach. They take advantage of the monotonicity properties of the option price with respect to the asset price and volatility, to obtain the conditional distribution of what they call the 'true' BS price as well as the conditional distribution of what they call the 'predicted' BS price. The former is obtained by conditioning on volatility (they assume that volatility is known and not estimated), while the latter follows from conditioning on the underlying asset price and treating the only source of randomness as being due to the classical variance estimate. In our analysis we go one step further, since we are able to account simultaneously for randomness in price and volatility.

[2]The term 'flexible' indicates the fact that different point or interval estimators can be obtained from the distribution of the option price.

[3]As has been suggested by Black (1976), Epps and Epps (1976), Morgan (1976), Christie (1982) and Tauchen and Pitts (1983).

As a Bayesian problem, the randomness of the BS price is unusual in that it depends both on data (price) and parameters (volatility). Under the model's assumptions (i.e. log-normality of stock prices), we derive the prior and posterior densities for a European call. We also provide expressions for the density of a call option conditional on the sample estimate of volatility and conditional on the asset price respectively. To this end, we investigate how the dispersion of the option price changes as we condition on more information – from the prior density, to conditioning only on the sample variance, to conditioning on the price, to the posterior density. The results we present, for a number of realistic values, show the extent to which conditioning on the asset price dramatically reduces the variability of the option price. Turning to the posterior distribution, we find that it exhibits excess kurtosis and is positively skewed, particularly so as we move progressively away from the money. This will have important implications for value-at-risk calculations, since such calculations critically depend on the left tail of the distribution under consideration.

Our chapter could be criticized on the grounds that option pricing has moved a long way from the BS model. Our response is that the BS model is still widely used in applications, especially in real options. For a detailed list of applications (e.g. in real options, bankruptcy problems, evaluation of insured bank deposits, actuarial work, etc.), see Knight and Satchell (1997). Moreover, recent empirical evidence (see Dumas *et al.*, 1998) suggests in favour of building on the deficiencies of Black–Scholes rather than abolishing it. In particular, these authors show that the predictive and hedging performance of the (deterministic) time-varying volatility models proposed by Derman and Kani (1994), Dupire (1994) and Rubinstein (1994) is no better than an *ad hoc* procedure that merely smooths Black–Scholes implied volatilities across exercise prices and times to expiration.

The organization of the chapter is as follows. In Section 6.1 we present the main theory. Subsection 6.1.1 introduces the stochastic model assumed to generate the data and sets up a Bayesian framework. In subsections 6.1.2 and 6.1.3 we work towards deriving the prior and posterior densities of the option price. Section 6.2 deals with their numerical evaluation. In Section 6.3 we present our results, and Section 6.4 concludes.

6.2 Derivation of the prior and posterior densities

6.2.1 Distributional assumptions

In parametric derivative pricing models, such as the BS, the price process of the underlying asset is fully specified up to a finite number of unknown parameters.[4] Here we use the traditional log-normal diffusion with unknown drift and volatility. It is therefore assumed that in the continuous-time limit the asset price at time t is P_t where P_t is determined by the stochastic differential equation:

$$dP_t = \mu P_t dt + \sigma P_t dW_t \tag{6.1}$$

[4]This is in contrast to non-parametric derivative pricing models, where the price process is not explicitly specified but is rather inferred from the data under suitable regularity conditions (see, for example, Hutchinson *et al.*, 1994; Jackwerth and Rubinstein, 1996).

where μ is the expected rate of return, σ is the the volatility, and $\{W_t, t \geq 0\}$ is standard Brownian motion. Then the asset price process may be represented as:

$$P_t = P_0 \exp\left[\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma(W_t - W_0)\right] \tag{6.2}$$

The geometric (continuously compounded) return between time 0 and t is:

$$x_t = \log(P_t/P_0) \tag{6.3}$$

From equation (6.1) it follows that the log-return x_t is generated by an independent normal process with $x_t \sim N((\mu - \sigma^2/2)t, \sigma^2 t)$.[5]

This then implies that $\log P_t \sim N(\log P_0 + (\mu - \frac{1}{2}\sigma^2)\, t, \sigma^2 t)$.

Assumption 1: *The asset price P_t is log-normally distributed. Its conditional probability density function is given by:*[6]

$$pdf(P_t \backslash P_0, \mu, \sigma, t) = \frac{1}{P_t\sqrt{2\pi t}\sigma} \exp\left\{\frac{-\left[\ln\left(\frac{P_t}{P_0}\right) - \left(\mu - \frac{\sigma^2}{2}\right)t\right]^2}{2\sigma^2 t}\right\} \quad \begin{array}{l} 0 \prec P_t \prec \infty \\ -\infty \prec \mu \prec \infty. \\ \sigma \succ 0 \end{array}$$

Assumption 2: *When the variance σ^2 of an independent normal process is assumed known but the mean μ is a random variable, the most convenient distribution for μ (the natural conjugate of the likelihood of the sample) is the normal distribution. The conditional probability density function of the expected rate of return μ is therefore given by:*

$$pdf(\mu \backslash \sigma, m, t) = \frac{\sqrt{t}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{t(\mu - m)^2}{2\sigma^2}\right) \begin{array}{l} -\infty \prec \mu, m \prec \infty, \\ \sigma \geq 0, t \succ 0 \end{array}.$$

where m is a hyperparameter.

Since the log-return x between two consecutive time intervals is normally distributed with variance σ^2, the classical minimum-variance unbiased estimator of σ^2 for t observations is

$$s^2 = \sum_{i=1}^{t} (x_i - \bar{x})^2/(t-1) \tag{6.4}$$

where $\bar{x} = (1/t)\sum_{j=1}^{t} x_j$. It is well known that the statistic $(t-1)s^2/\sigma^2$ has a χ^2 distribution with $t-1$ degrees of freedom. This in turn implies that the estimator s^2 is distributed $\sigma^2\chi^2/(t-1)$ with $t-1$ degrees of freedom.

[5] $N(\ )$ denotes the normal distribution.

[6] Here and below we use *pdf* to denote probability density functions generally, and not one specific probability density. The argument of *pdf* as well as the context in which it is used will identify the particular *pdf* being considered.

Assumption 3: *The likelihood function for σ^2 is therefore defined as:*[7]

$$L(\sigma^2 \backslash s^2, t) \equiv f_{\chi 2}\left(\frac{(t-1)s^2}{\sigma^2}\right)\frac{t-1}{\sigma^2}$$

$$\Rightarrow pdf(s^2 \backslash \sigma^2, t) = \left(\frac{t-1}{2}\right)^{\frac{t-1}{2}} \frac{(s^2)^{\frac{(t-1)}{2}-1}}{\Gamma\left(\frac{t-1}{2}\right)(\sigma^2)^{\frac{t-1}{2}}} \exp\left(-\frac{(t-1)s^2}{2\sigma^2}\right).$$

$$\Rightarrow pdf(s \backslash \sigma, \upsilon) = \frac{2\left(\frac{\upsilon}{2}\right)^{\frac{\upsilon}{2}}}{\Gamma\left(\frac{\upsilon}{2}\right)} \frac{s^{\upsilon-1}}{\sigma^{\upsilon}} \exp\left(-\frac{\upsilon s^2}{2\sigma^2}\right)$$

where $\upsilon = t - 1$.

The fact that $(t-1)s^2/\sigma^2 \sim \chi^2_{(t-1)}$ suggests that the conditional probability density function for the variance (i.e. $pdf(\sigma^2 \backslash s^2, t)$) is inverted-gamma-1. We use this as motivation when choosing the prior density for σ^2.

Assumption 4: *We can assign an inverted-gamma-1 distribution with hyperparameters λ, θ as the prior distribution for σ^2. Its prior probability density function is then given by:*[8]

$$pdf(\sigma^2 \backslash \lambda, \theta) = f_{i\gamma}(\sigma^2 \backslash \lambda, \theta) = \frac{\lambda^{\theta}}{\Gamma(\theta)} \frac{1}{(\sigma^2)^{\theta+1}} \exp\left(-\frac{\lambda}{\sigma^2}\right)$$

$$\Rightarrow pdf(\sigma \backslash \lambda, \theta) = 2\sigma \frac{\lambda^{\theta}}{\Gamma(\theta)} \frac{1}{(\sigma^2)^{\theta+1}} \exp\left(-\frac{\lambda}{\sigma^2}\right) \quad \lambda, \theta \succ 0.$$

A Bayesian framework has now been introduced. Assumptions 2 and 4 define prior distributions for the process parameters μ and σ. Assumption 3 defines the likelihood for the variance (volatility). The process for the stock price is represented in equation (6.2), while the conditional probability density function for the stock price is defined in Assumption 1.

Karolyi (1993) provided a Bayesian analysis for the stock return volatility. He combined an inverted-gamma prior density (i.e. $pdf(\sigma^2 \backslash \vartheta)$)[9] together with the likelihood (i.e. $L(\sigma^2 \backslash s^2, t)$) to obtain the posterior probability density function of the variance (i.e. $pdf(\sigma^2 \backslash s^2, t; \vartheta)$). We extend his analysis by deriving the posterior probability

[7] $f_{\chi 2}(\,)$ denotes the χ^2 probability density function.
[8] $f_{i\gamma}(\,)$ denotes the inverted-gamma-1 probability density function.
[9] ϑ is a two-dimensional vector of prior parameters estimated using information extracted from the cross-sectional patterns in return volatilities for groups of stocks sorted on size, financial leverage and trading volume.

distribution of the option price, not just as a transformation of the posterior density of the variance but by incorporating randomness in the underlying asset price as well.

Remark 1:

- Our Bayesian framework has been established on the basis of geometric Brownian motion for the stock price process. It is just as easy to establish a Bayesian framework for an Ornstein–Uhlenbeck (O–U) price process: $(dP_t = -\phi(P_t - \mu)dt + \sigma dW_t)$, since it is also a normal Markovian process with stationary transition probabilities.
- It can be shown (see, for example, Nelson, 1990) that typical discrete-time models of heteroscedasticity, including certain ARCH and EGARCH models, converge in a natural way as the time intervals shrink to the continuous-time stochastic volatility model in which the log of volatility is well defined and satisfies the Ornstein–Uhlenbeck differential equation. Thus an interesting extension of this chapter could be a Bayesian analysis of a stochastic volatility model.

6.2.2 The prior density

For $0 \leq t \leq T$, the time t price of a European call option at strike price K with expiry time $\tau = T - t$ is:

$$c = C(P_t, K, \sigma, \tau, r) \tag{6.5}$$
$$= P_t \Phi\left(\frac{\log\left[\frac{P_t}{K}\right] + \left[r + \frac{1}{2}\sigma^2\right]\tau}{\sigma\sqrt{\tau}}\right) - K\exp(-r\tau)\Phi\left(\frac{\log\left[\frac{P_t}{K}\right] + \left[r - \frac{1}{2}\sigma^2\right]\tau}{\sigma\sqrt{\tau}}\right)$$

where r is the risk-free interest rate (assumed fixed and known from 0 to T) and everything else as already defined. The term $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. We shall either condition on time t or on time 0 information, and the latter we shall refer to as unconditional.

Remark 2: The information available at time t is the history of the discrete price process, $(P_0, P_1, P_2, \ldots, P_t)$. Then conditionally on time t, randomness in the option price stems only from the unknown volatility σ. One may therefore write $c = C(\sigma)$.

As aforementioned, literature on the estimation of σ^2 abounds. The benchmark procedure is to use the classical estimator s^2 given in equation (6.4). The paper by Boyle and Ananthanarayanan (1977) first evaluates the impact of variance estimation in option valuation models. The authors recognize that using s^2 as an estimate of the variance does produce biased option prices.[10] However, they claim that the magnitude of the bias is

[10]This is because of the non-linearity of the option price formula.

not large and they are more concerned with the dispersion induced in the option price. Interestingly, they suggest that a Bayesian approach may be usefully employed to improve on the precision of option price estimates.

Butler and Schachter (1986), on the other hand, are concerned with the variance-induced option price bias, and investigate potential remedial measures. In fact they construct a uniformly minimum-variance unbiased estimator for the BS option price. The estimator is derived by taking a Taylor series expansion of the pricing formula and the moments of the estimated variance.

In a discussion of the Butler and Schachter paper, Knight and Satchell (1997) re-examine the question of statistical bias in the BS option price. They show that the only unbiased estimated option is an at-the-money option. However, they argue that the importance of bias in option pricing seems minor compared with other obvious sources of mispricing.

Noh *et al.* (1994) assess the performance of ARCH models for pricing options. Rather than comparing implied volatilities with GARCH volatilities, their study compares predictions of options prices from GARCH with predictions of the same options prices from forecasting implied volatility. The results indicate that both methods can effectively forecast prices well enough to profit by trading if transaction costs are not too high. The GARCH models are considerably more effective.

Remark 3: Unconditionally, the option price depends both on the volatility σ and the stock price process $P_t = P_0 \exp[(\mu - (1/2)\sigma^2)t + \sigma(W_t - W_0)]$. We write $c = C(P_t, \sigma)$ to denote that fact. Bayesian theory mandates that this should be taken into account when deriving the posterior density. In other words we should treat prices and volatility as unknown random variables and identify a prior density for them.

The building block for the derivation of the prior density of the BS option price[11] is therefore the joint density function of P_t and σ; i.e. $pdf(P_t, \sigma \backslash P_0, \lambda, \theta, t)$. Then, by transforming $pdf(P_t, \sigma \backslash P_0, \lambda, \theta, t)$ to $pdf(c, \sigma \backslash P_0, \lambda, \theta, t)$ and integrating out σ, we obtain $pdf(c \backslash P_0, \lambda, \theta, t)$. Note that when the distributions are conditional on any prior parameters (i.e. λ, θ and m), and/or on P_0, and/or on t (it is not unreasonable to assume that the sample size is known before the sample is drawn), we will refer to these distributions as prior or unconditional. In what follows, for ease of notation, we shall ignore the dependence on P_0, λ, θ and t.

Proposition 1: The joint unconditional density of P_t and σ is given by

$$pdf(P_t, \sigma) = \frac{1}{\sqrt{\pi t} P_t} \frac{\lambda^\theta}{\Gamma(\theta)} \frac{1}{\sigma^{2(\theta+1)}} \exp\left(-\frac{\lambda}{\sigma^2}\right) \times \exp\left(-\frac{1}{4\sigma^2 t}\left[\ln\left(\frac{P_t}{P_0}\right) - \left(m - \frac{1}{2}\sigma^2\right)t\right]^2\right)$$

[11] When we refer to the probability density of the BS option price, we mean the probability density of a European call option.

Proof: From Assumptions 4 and 2 we have $pdf(\sigma)$ and $pdf(\mu\backslash\sigma)$ respectively. $pdf(\mu, \sigma) = pdf(\sigma)pdf(\mu\backslash\sigma)$. Similarly, $pdf(P_t, \mu, \sigma) = pdf(\mu, \sigma)pdf(P_t\backslash\mu, \sigma)$ with $pdf(P_t\backslash\mu, \sigma)$ given in Assumption 1. Finally, $pdf(P_t, \sigma) = \int_{-\infty}^{\infty} pdf(P_t, \mu, \sigma)\partial\mu$. Section A.1 in the Appendix contains the analytic proof and all the relevant calculations.

We are now in a position to derive the unconditional density function of the option price. Let us first obtain $pdf(c, \sigma)$: Take $pdf(P_t, \sigma)$ and consider the transformation

$$c = C(P_t) \equiv C(P_t, \sigma) \Leftrightarrow P_t = \Psi(c) \equiv \Psi(c, \sigma)$$
$$\sigma = \sigma$$

where Ψ is the inverse of the option price with respect to P_t.[12] The Jacobian J of the transformation is given by:

$$\frac{1}{J} = \begin{vmatrix} \dfrac{\partial c}{\partial P_t} = \Phi(d_1) & \dfrac{\partial \mathrm{c}}{\partial \sigma} = vega \\ \dfrac{\partial \sigma}{\partial P_t} = 0 & \dfrac{\partial \sigma}{\partial \sigma} = 1 \end{vmatrix} = \Phi(d_1) \tag{6.6}$$

where $d_1 = \left[\ln\left(\frac{P_t}{Ke^{-r\tau}}\right) + \frac{\sigma^2\tau}{2}\right] / \sigma\sqrt{\tau}$ and $vega = \phi\left(\frac{\ln\left(\frac{P_t}{Ke^{-r\tau}}\right) + \frac{\sigma^2\tau}{2}}{\sigma\sqrt{\tau}}\right) P_t\sqrt{\tau}$, and $\phi(\ldots) = \Phi'(\ldots)$ denotes the standard normal probability density function. Then

$$\begin{aligned} pdf(c, \sigma) &= pdf(\Psi(c, \sigma), \sigma)\,|J| \\ &= \frac{1}{\Phi(d_1^*)\sqrt{\pi t}\Psi(c, \sigma)} \frac{\lambda^\theta}{\Gamma(\theta)} \frac{1}{\sigma^{2(\theta+1)}} \\ &\quad \times \exp\left(-\frac{\lambda}{\sigma^2} - \frac{1}{4\sigma^2 t}\left[\ln\left(\frac{\Psi(c, \sigma)}{P_0}\right) - \left(m - \frac{1}{2}\sigma^2\right)t\right]^2\right) \end{aligned} \tag{6.7}$$

where $d_1^* = \left[\ln\left(\frac{\Psi(c, \sigma)}{Ke^{-r\tau}}\right) + \frac{\sigma^2\tau}{2}\right] / \sigma\sqrt{\tau}$.

[12] We invert the option pricing formula in terms of P_t, hence obtaining P_t as a function of c and σ. It should, however, be noted that there is no analytic expression (with the exception of an at-the-money option) for $P_t = \Psi(c, \sigma)$ and a Newton–Raphson numerical approximation is required.

Integrating out σ will give us the prior density of the option price:

$$pdf(c) = \int_0^\infty pdf(c, \sigma)\partial\sigma \tag{6.8}$$

Note, however, that there is no closed form solution for this integral and it will have to be evaluated numerically (more of that in Section 6.2).

Remark 4: We can utilize another procedure to obtain the marginal density of the option price $pdf(c)$. We start again from $pdf(P_t, \sigma)$ but this time we consider the transformation

$$c = C(\sigma) \equiv C(P_t, \sigma) \Leftrightarrow \sigma = \Theta(c) \equiv \Theta(c, P_t)$$
$$P_t = P_t$$

where Θ is the inverse of the option price with respect to σ.[13] This is commonly referred to as the implied volatility of the option price. The Jacobian J of the transformation is given by

$$\frac{1}{J} = \begin{vmatrix} \dfrac{\partial c}{\partial \sigma} = vega & \dfrac{\partial c}{\partial P_t} = \Phi(d_1) \\ \dfrac{\partial P_t}{\partial \sigma} = 0 & \dfrac{\partial P_t}{\partial P_t} = 1 \end{vmatrix} = vega$$

Thus we obtain:

$$\begin{aligned} pdf(c, P_t) &= pdf(\Theta(c, P_t), P_t)\,|J'| \\ &= \frac{|J'|}{\sqrt{\pi t} P_t} \frac{\lambda^\theta}{\Gamma(\theta)} \frac{1}{\Theta(c, P_t)^{2(\theta+1)}} \\ &\quad \times \exp\left(-\frac{\lambda}{\Theta(c, P_t)^2} - \frac{1}{4\sigma^2 t} \left[\ln\left(\frac{P_t}{P_0}\right) - \left(m - \frac{1}{2}\Theta(c, P_t)^2 \right) t \right]^2 \right) \end{aligned}$$

where $J' = 1 \Big/ \phi\left(\dfrac{\left[\ln\left(\dfrac{P_t}{Ke^{-r\tau}}\right) + \dfrac{\Theta(c, P_t)^2 \tau}{2} \right]}{\Theta(c, P_t)\sqrt{\tau}} \right) P_t \sqrt{\tau}.$

[13] We invert the option pricing formula in terms of σ, hence obtaining σ as a function of c and P_t. Note again that there is no analytic expression (with the exception of an at-the-money option) for $\sigma = \Theta(P_t, c)$ and a Newton–Raphson numerical approximation is required.

Integrating out P_t gives us the prior density of the option price:

$$pdf(c) = \int_0^\infty pdf(c, P_t)\partial P_t$$

Needless to say, either procedure gives the same (numerical) values for $pdf(c)$. Unless otherwise stated, in our later calculations we shall use the first procedure, given by equation (6.8).

6.2.3 The posterior density

In contrast to classical analysis, where the main piece of output is a point estimate, Bayesian analysis produces as its main piece of output the so-called posterior density. This posterior density can then be combined with a loss or utility function to allow a decision to be made on the basis of minimizing expected loss or maximizing expected utility. For example, for positive definite quadratic loss functions the mean of the posterior distribution is an optimal point estimate. If the loss is proportional to the absolute value of the difference between the true and the estimated values, the median is chosen, while a zero loss for a correct estimate and a constant loss for an incorrect estimate lead to the choice of the mode.

To derive the posterior density of the option price, which unconditionally depends on two stochastic arguments (namely the price process and volatility) while conditionally only on volatility, we proceed as follows.

By Bayes rule:

$$pdf(P_t, \sigma \backslash s) = \frac{pdf(P_t, \sigma)pdf(s \backslash \sigma, P_t)}{pdf(s)} = \frac{pdf(P_t, \sigma)pdf(s \backslash \sigma)}{pdf(s)} = \frac{pdf(P_t, \sigma, s)}{pdf(s)} \tag{6.9}$$

We therefore require expressions for $pdf(P_t, \sigma, s)$ and $pdf(s)$.

Proposition 2:

$$pdf(P_t, \sigma, s) = \frac{2}{\sqrt{\pi t} P_t} \frac{\lambda^\theta}{\Gamma\left(\frac{\upsilon}{2}\right)\Gamma(\theta)} \left(\frac{\upsilon}{2}\right)^{\frac{\upsilon}{2}} s^{\upsilon-1} \frac{1}{\sigma^{2\theta+\upsilon+2}} \times \exp\left(-\frac{2\lambda + \upsilon s^2}{2\sigma 2} - \frac{1}{4\sigma^2 t}\left[\ln\left(\frac{P_t}{P_0}\right) - \left(m - \frac{1}{2}\sigma^2\right)t\right]^2\right)$$

Proof: From Proposition 1 we have $pdf(P_t, \sigma)$. Also, from Assumption 3 we have $pdf(s \backslash \sigma)[= pdf(s \backslash \sigma, P_t)]$. Then $pdf(P_t, \sigma, s) = pdf(P_t, \sigma)pdf(s \backslash \sigma)$ and the result follows.

Proposition 3: The unconditional density of the statistic s is given by

$$pdf(s) = \frac{2}{\mathrm{B}\left(\frac{\upsilon}{2}, \theta\right)} \lambda^{\theta} \left(\frac{\upsilon}{2}\right)^{\frac{\upsilon}{2}} \frac{s^{\upsilon-1}}{\left(\lambda + \frac{\upsilon s^2}{2}\right)^{\frac{\upsilon}{2}+\theta}}$$

Proof: From Assumptions 4 and 3, we know $pdf(\sigma^2)$ and $pdf(s^2 \backslash \sigma^2)$. Then $pdf(s^2, \sigma^2) = pdf(\sigma^2) pdf(s^2 \backslash \sigma^2)$. We can now integrate out σ^2 to obtain the marginal probability density of s^2, i.e. $pdf(s^2) = \int_0^{\infty} pdf(s^2, \sigma^2) \partial\sigma^2$. Finally, $pdf(s) = 2s pdf(s^2)$. The complete proof and all the relevant calculations are presented in Section A.2 in the Appendix.

Now take $pdf(P_t, \sigma, s)$ and consider the transformation:

$$\begin{aligned} c &= C(\sigma) \equiv C(P_t, \sigma) \Leftrightarrow \sigma = \Theta(c) \equiv \Theta(c, P_t) \\ P_t &= P_t \\ s &= s \end{aligned}$$

where Θ is the inverse of the option price with respect to σ. The Jacobian of the transformation is given by:

$$\frac{1}{J} = \begin{vmatrix} \partial c/\partial\sigma = vega & \partial c/\partial P_t = \Phi(d_1) & \partial c/\partial s = 0 \\ \partial P_t/\partial\sigma = 0 & \partial P_t/\partial P_t = 1 & \partial P_t/\partial s = 0 \\ \partial s/\partial\sigma = 0 & \partial s/\partial P_t = 0 & \partial s/\partial s = 1 \end{vmatrix} = vega \tag{6.10}$$

Then

$$pdf(P_t, c, s) = pdf(P_t, \Theta(c, P_t), s)\,|J| \tag{6.11}$$

We can now obtain an expression for $pdf(P_t, c \backslash s)$:

$$\begin{aligned} pdf(P_t, c \backslash s) &= \frac{pdf(P_t, c, s)}{pdf(s)} \qquad (6.12) \\ &= \frac{|J^*|}{\sqrt{\pi t} P_t \Gamma\left(\frac{\upsilon}{2} + \theta\right)} \frac{\left(\lambda + \frac{\upsilon s^2}{2}\right)^{\frac{\upsilon}{2}+\theta}}{\Theta(c, P_t)^{2\theta+\upsilon+2}} \\ &\quad \times \exp\left(-\frac{2\lambda + \upsilon s^2}{2\Theta(c, P_t)^2} - \frac{1}{4\Theta(c, P_t)^2 t}\left[\ln\left(\frac{P_t}{P_0}\right) - \left(m - \frac{1}{2}\Theta(c, P_t)^2\right) t\right]^2\right) \end{aligned}$$

where $J^* = 1 \Big/ \left(\phi \left(\dfrac{\ln\left(\dfrac{P_t}{Ke^{-r\tau}}\right) + \dfrac{\Theta(c, P_t)^2 \tau}{2}}{\Theta(c, P_t)\sqrt{\tau}} \right) P_t \sqrt{\tau} \right)$.

Having obtained $pdf(P_t, c\backslash s)$; the posterior density of the option price is given by:

$$pdf(c\backslash P_t, s) = \frac{pdf(P_t, c\backslash s)}{pdf(P_t\backslash s)} \tag{6.13}$$

We derived $pdf(P_t, c\backslash s)$ in equation (6.12), but we need an expression for $pdf(P_t\backslash s)$.

Proposition 4: The conditional (on the sample and prior information) density for the asset price is given by

$$pdf(P_t\backslash s) = \frac{K_{\theta+\frac{t}{2}}\left(\frac{1}{4}\sqrt{2(2\lambda+\upsilon s^2)t + (\ln(P_t/P_0) - mt)^2}\right)}{\sqrt{\pi t} P_t \Gamma\left(\theta + \frac{\upsilon}{2}\right)}$$
$$\times \left(\lambda + \frac{\upsilon}{2}s^2\right)^{\theta+\frac{\upsilon}{2}} \left(\frac{t}{16}\right)^{\theta+\frac{t}{2}} \left[\frac{2(2\lambda+\upsilon s^2)t + (\ln(P_t/P_0) - mt)^2}{64}\right]^{\frac{2\theta+t}{4}}$$
$$\times \exp\left(-\frac{\ln(P_t/P_0) - mt}{4}\right)$$

where $K_{\theta+\frac{t}{2}}\left(\frac{1}{4}\sqrt{2(2\lambda t+\upsilon s^2)t + (\ln(P_t/P_0) - mt)^2}\right)$ is the modified Bessel function[14] of the second kind of order $\theta + (t/2)$.

[14]Modified Bessel functions are solutions to the differential equation $x^2y'' + xy' - (x^2 + a^2)y = 0$.

Definition 2: *Bessel Functions*

The differential equation $x^2y'' + xy' + (x^2 - a^2)y = 0$ is known as the Bessel equation where a is a non-negative constant. Some of its solutions are known as *Bessel functions*. The function J_a defined by $J_a(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!(n+a)!}\left(\frac{x}{2}\right)^{2n+a}$ for $x > 0$ and a a non-negative integer is called the *Bessel function of the first kind of order a*. The function K_a defined for $x > 0$ by

$$K_a(x) = J_a(x)\ln x - \frac{1}{2}\left(\frac{x}{2}\right)^{-a}\sum_{n=0}^{a-1}\frac{(a-n-1)!}{n!}\left(\frac{x}{2}\right)^{2n} - \frac{1}{2}\left(\frac{x}{2}\right)^{a}\sum_{n=0}^{\infty}(-1)^n\frac{h_n + h_{n+a}}{n!(n+a)!}\left(\frac{x}{2}\right)^{2n}$$

is called the *Bessel function of the second kind of order a*. The general solution of the Bessel equation in this case for $x > 0$ is $y = c_1J_a(x) + c_2K_a(x)$. For an exposition of Bessel functions and their relevance in diffusion theory, see, for example, Feller (1971). For a detailed discussion, see Watson (1944).

Proof: Let us first illustrate how to obtain the marginal density of the asset price – i.e. $pdf(P_t)$. Then it is straightforward to also obtain $pdf(P_t \backslash s)$ using a similar procedure. In Proposition 1 we have obtained $pdf(P_t, \sigma)$, which we can straightforwardly transform to $pdf(P_t, \sigma^2)$. Then the result follows from the fact that $pdf(P_t) = \int_0^\infty pdf(P_t, \sigma^2)\partial\sigma^2$ can be written as $A\int_0^\infty \frac{c^\omega}{\Gamma(\omega)}(\frac{1}{\sigma^2})^{\omega+1}\exp(-\frac{c}{\sigma^2})\exp(-p\sigma^2)\partial\sigma^2$ where A is a constant. It is easy now to observe that the integral is the *Laplace transform* of an inverted-gamma function ($L(f_{i\gamma})$): $L(f_{i\gamma}) = F(p) = \int_0^\infty f_{i\gamma}(\sigma^2)e^{-p\sigma^2}d\sigma^2$. For the complete proof, see Section A.3 in the Appendix.

We have now completed the derivation of the posterior density of the BS option price. Let us present here the full expression:

$$
\begin{aligned}
pdf(c\backslash P_t, s) = & \frac{|J^*|\left(\frac{2(2\lambda+\upsilon s^2)t+(\ln(P_t/P_0)-mt)^2}{64}\right)^{\frac{2\theta+t}{4}}}{K_{\theta+\frac{t}{2}}\left(\frac{1}{4}\sqrt{2(2\lambda+\upsilon s^2)t+(\ln(P_t/P_0)-mt)^2}\right)\Theta(c,P_t)^{2\theta+\upsilon+2}\left(\frac{t}{16}\right)^{\theta+\frac{t}{2}}} \\
& \times\exp\left(-\frac{2\lambda+\upsilon s^2}{2\Theta(c,P_t)^2}+\frac{\ln(P_t/P_0)-mt}{4}\right) \\
& \times\exp\left(-\frac{1}{4\Theta(c,P_t)^2 t}\left[\ln(P_t/P_0))-\left(m-\frac{1}{2}\Theta(c,P_t)^2\right)t\right]^2\right) \qquad (6.14)
\end{aligned}
$$

where $J^* = 1\Big/\left(\phi\left(\frac{\ln\left(\frac{P_t}{Ke^{-r\tau}}\right)+\frac{\Theta(c,P_t)^2\tau}{2}}{\Theta(c,P_t)\sqrt{\tau}}\right)P_t\sqrt{\tau}\right)$.

Corollary 1: *If the option is at-the-money, i.e.* $P_t = K\exp(-r\tau)$, *then certain simplifications occur:* $\Theta(c,P_t) = \frac{2}{\sqrt{\tau}}\Phi^{-1}(\frac{1}{2}(\frac{c}{P_t}+1))$ *and* $J^* = 1/(\phi(\Phi^{-1}(\frac{1}{2}(\frac{c}{P_t}+1)))P_t\sqrt{\tau})$.

Proof: When the option is at-the-money, the BS formula (given in equation (6.5)) simplifies to $c = C(P_t = K\exp(-r\tau), \sigma) = P_t(\Phi(\frac{\sigma\sqrt{\tau}}{2}) - \Phi(-\frac{\sigma\sqrt{\tau}}{2})) = P_t(2\Phi(\frac{\sigma\sqrt{\tau}}{2})-1)$. This then implies that: $\Phi(\frac{\sigma\sqrt{\tau}}{2}) = \frac{1}{2}(\frac{c}{P_t}+1) \Rightarrow \sigma = \Theta(c,P_t) = \frac{2}{\sqrt{\tau}}\Phi^{-1}(\frac{1}{2}(\frac{c}{P_t}+1))$ where $\Phi^{-1}(\ldots)$ denotes the inverse cumulative normal distribution function.

Also for $P_t = K\exp(-r\tau)$ we have $J^* = 1\Big/\left(\phi\left(\frac{\Theta(c,P_t)\sqrt{\tau}}{2}\right)P_t\sqrt{\tau}\right)$. Substituting in, the analytic expression for $\Theta(c,P_t)$ we get the proposed result for J^*.

Remark 5: The at-the-money case is best interpreted as a stochastic exercise price where $K = P_t\exp(r\tau)$.

Remark 6: It is interesting to observe that the posterior density of the option price does depend on the expected rate of return μ through the hyperparameter m (m represents our prior beliefs about μ). The true unknown μ has been integrated out. The existence of m in the formula is due to randomness in prices prior to sampling.

Corollary 2: *In relation to the above remark, market completeness implies that we can also derive the risk-neutral posterior density of the option price. Under the risk-neutral measure the variance of the price process remains the same but the drift changes and is equal to the risk-free rate r. Under the assumption of a constant risk-free rate this effectively implies that in Assumption 1 we substitute $r = m$ and also Assumption 2 is no longer required since r is constant and not a random variable. Following the same procedure as above (without Assumption 2 this time), we can therefore show that the risk-neutral posterior density is given by:*

$$pdf^{RN}(c\backslash P_t, s) = \frac{|J^*| \left(\frac{(2\lambda + \upsilon s^2)t + (\ln(P_t/P_0) - rt)^2}{16} \right)^{\frac{2\theta+t}{4}}}{K_{\theta+\frac{t}{2}} \left(\frac{1}{2}\sqrt{(2\lambda + \upsilon s^2)t + (\ln(P_t/P_0) - rt)^2} \right) \Theta(c, P_t)^{2\theta+\upsilon+2} \left(\frac{t}{8} \right)^{\theta+\frac{t}{2}}}$$
$$\times \exp\left(-\frac{2\lambda + \upsilon s^2}{2\Theta(c, P_t)^2} + \frac{\ln(P_t/P_0) - rt}{2} \right)$$
$$\times \exp\left(-\frac{1}{2\Theta(c, P_t)^2 t} \left[\ln(P_t/P_0) - \left(r - \frac{1}{2}\Theta(c, P_t)^2 \right) t \right]^2 \right)$$

Note, however, that this is as far as we choose to go with the risk-neutral measure. In the context of the problem we examine, it does not make sense to derive a risk-neutral predictive density of the unconditional price process. In other words the predictive density is not invariant in m. The same applies, of course, to the prior density as well.

Having derived the prior and posterior densities of the BS option price (i.e. $pdf(c)$ and $pdf(c\backslash P_t, s)$), it is interesting, for comparative purposes in particular, to derive expressions for $pdf(c\backslash s)$ and $pdf(c\backslash P_t)$. This way we can illustrate how the dispersion of the density of the option price changes as we condition on more information: from the prior $pdf(c)$, to conditioning only on the sample estimate of volatility $pdf(c\backslash s)$, to conditioning on the price $pdf(c\backslash P_t)$, to the posterior density $pdf(c\backslash P_t, s)$. We refer the reader to Section A.4 in the Appendix for the derivation of $pdf(c\backslash s)$ and $pdf(c\backslash P_t)$.

It should be stressed that randomness in prices and volatility has been assumed throughout our analysis. We write $c = C(P_t, \sigma)$ to denote that fact. For fixed prices but random volatility, we would write $c = C(\sigma)$. Note for example that $pdf(c = C(P_t, \sigma)\backslash s)$ and $pdf(c = C(\sigma)\backslash s)$ represent two very different densities with dramatically different shapes.[15]

[15]The former represents the density of the option price conditional on the sample estimate of volatility but with prices unknown, while the latter represents the density of the option price conditional on the sample estimate of volatility but with prices known and fixed. Thus the dispersion of the former distribution is expected to be much larger than the latter.

Remark 7: So far it has been assumed that $c = C(P_t, \sigma)$. Karolyi (1993) assumes that prices are non-random, i.e. $c = C(\sigma)$, and suggests that the posterior density of the option price can be derived as a non-linear transformation of the posterior density of volatility. Let us briefly illustrate how $pdf(c = C(\sigma)\backslash s)$ can be obtained. The posterior density of volatility is given by

$$pdf(\sigma\backslash s) = (pdf(\sigma)L(\sigma\backslash s)) \Big/ \int_0^\infty pdf(\sigma, s)\partial\sigma = \frac{pdf(\sigma, s)}{pdf(s)}$$

where $pdf(\sigma, s) = \dfrac{4\lambda^\theta}{\Gamma(\theta)\Gamma(\nu/2)} \left(\dfrac{\nu}{2}\right)^{\nu/2} \dfrac{s^{\nu-1}}{\sigma^{2\theta+\nu+1}} \exp\left(-\dfrac{2\lambda + \nu s^2}{2\sigma^2}\right)$

$$pdf(s) = \frac{2}{\mathrm{B}\left(\frac{\nu}{2}, \theta\right)} \lambda^\theta \left(\frac{\nu}{2}\right)^{\frac{\nu}{2}} \frac{s^{\nu-1}}{\left(\lambda + \frac{\nu s^2}{2}\right)^{\frac{\nu}{2}+\theta}}.$$

Then

$$pdf(\sigma\backslash s) = \frac{2}{\Gamma\left(\frac{\nu}{2} + \theta\right)} \frac{\left(\lambda + \frac{\nu s^2}{2}\right)^{\frac{\nu}{2}+\theta}}{\sigma^{2\theta+\nu+1}} \exp\left(-\frac{2\lambda + \nu s^2}{2\sigma^2}\right).$$

This is the posterior density of σ. Using the transformation $c = C(\sigma)$, i.e. inverting the Black–Scholes formula in terms of $\sigma = C^{-1}(c) = \Theta(c)$, we obtain $pdf(c = C(\sigma)\backslash s)$:

$$pdf(c = C(\sigma)\backslash s) = \frac{2\left(\lambda + \dfrac{\nu s^2}{2}\right)^{\frac{\nu}{2}+\theta} \exp\left(-\dfrac{2\lambda + \nu s^2}{2[\Theta(c)]^2}\right)}{\Gamma\left(\dfrac{\nu}{2} + \theta\right)[\Theta(c)]^{2\theta+\nu+1} \phi\left(\dfrac{\ln\left(\dfrac{P_t}{Ke^{-r\tau}}\right) + \dfrac{[\Theta(c)]^2\tau}{2}}{\Theta(c)\sqrt{\tau}}\right) P_t\sqrt{\tau}}$$

Again, for the at-the-money case, the simplifications outlined in Corollary 1 apply.

6.3 Numerical evaluation

In equation (6.7) we have derived the joint unconditional (prior) density of the option price and volatility:

$$pdf(c, \sigma) = \frac{1}{\Phi(d_1^*)\sqrt{\pi t}\Psi(c, \sigma)} \frac{\lambda^\theta}{\Gamma(\theta)} \frac{1}{\sigma^{2(\theta+1)}} \times \exp\left(-\frac{\lambda}{\sigma^2} - \frac{1}{4\sigma^2 t}\left[\ln\left(\frac{\Psi(c, \sigma)}{P_0}\right) - \left(m - \frac{1}{2}\sigma^2\right)t\right]^2\right)$$

$$d_1^* = \left[\ln\left(\frac{\Psi(c, \sigma)}{Ke^{-r\tau}}\right) + \frac{\sigma^2\tau}{2}\right] \Big/ \sigma\sqrt{\tau}$$

where $\Psi(c, \sigma)$ is the inverse of the option price c with respect to P_t (call it the implied price hereafter); λ, θ are the prior parameters of the volatility distribution; m is the prior expected rate of return of the asset; t is the sample size (it is reasonable to assume that the sample size is known although the sample is not yet drawn); τ is the time to maturity of the option under consideration; P_0, K and r are the initial asset price, the strike price, and the risk-free interest rate respectively; and $\Phi(.)$ is the cumulative standard normal distribution function.

To find the marginal (prior) density of the option price we need to integrate out the volatility parameter σ. However, this cannot be done analytically, and we will have to evaluate the density numerically.

Let us first specify our prior parameters; namely λ, θ and m. We have from Assumption 4 that:

$$pdf(\sigma) = 2\sigma \frac{\lambda^\theta}{\Gamma(\theta)} \frac{1}{(\sigma^2)^{\theta+1}} \exp\left(-\frac{\lambda}{\sigma^2}\right)$$

Taking the first and second moments of the distribution of volatility, we get

$$\begin{aligned} E(\sigma) &= \int_0^\infty \sigma \left[2\sigma \frac{\lambda^\theta}{\Gamma(\theta)} \frac{1}{(\sigma^2)^{\theta+1}} \exp\left(-\frac{\lambda}{\sigma^2}\right)\right] d\sigma \qquad (6.15) \\ &= \lambda^{1/2} \frac{\Gamma(\theta - 1/2)}{\Gamma(\theta)} \int_0^\infty 2\sigma \frac{\lambda^{\theta-1/2}}{\Gamma(\theta - 1/2)} \frac{1}{(\sigma^2)^{(\theta-\frac{1}{2})+1}} \exp\left(-\frac{\lambda}{\sigma^2}\right) d\sigma \\ &= \lambda^{1/2} \frac{\Gamma(\theta - 1/2)}{\Gamma(\theta)} \end{aligned}$$

$$\begin{aligned} E(\sigma^2) &= \int_0^\infty \sigma^2 \left[2\sigma \frac{\lambda^\theta}{\Gamma(\theta)} \frac{1}{(\sigma^2)^{\theta+1}} \exp\left(-\frac{\lambda}{\sigma^2}\right)\right] d\sigma \qquad (6.16) \\ &= \frac{\lambda}{\theta - 1} \int_0^\infty 2\sigma \frac{\lambda^{\theta-1}}{\Gamma(\theta - 1)} \frac{1}{(\sigma^2)^{(\theta-1)+1}} \exp\left(-\frac{\lambda}{\sigma^2}\right) d\sigma \\ &= \frac{\lambda}{\theta - 1} \end{aligned}$$

Also,

$$Var(\sigma) = E(\sigma^2) - [E(\sigma)]^2 \Rightarrow E(\sigma^2) = Var(\sigma) + [E(\sigma)]^2 \qquad (6.17)$$

Once we have prior beliefs about the mean and variance of volatility (i.e. $E(\sigma)$ and $Var(\sigma)$), we can calculate λ, θ using equations (6.15), (6.16) and (6.17) above. Our prior beliefs, i.e. $E(\mu)$, will also determine the value of m. We digress briefly to discuss how λ, θ might be chosen.

One version is the 'empirical' Bayes approach. A prior is constructed from the data themselves, and so can be viewed as incorporating a non-informative prior. The Stein estimator can be viewed as an empirical Bayes estimator (see Efron and Morris, 1973). Prior sample data could also act as a useful source of information when forming prior

beliefs. This is because of the clustering effect: observations of financial time series reveal bunching of high- and low-volatility episodes. Alternatively, one could use the long run average of volatility as prior information to capture the mean reverting behaviour of volatility.[16]

We want to solve the Black–Scholes equation in terms of P_t and thus obtain $P_t = \Psi(c, \sigma)$ (i.e. obtain the price of the underlying as a function of the option price and of volatility). But

$$c = P_t \Phi\left(\frac{\log\left[\frac{P_t}{K}\right] + \left[r + \frac{1}{2}\sigma^2\right]\tau}{\sigma\sqrt{\tau}}\right) - K\exp(-r\tau)\Phi\left(\frac{\log\left[\frac{P_t}{K}\right] + \left[r - \frac{1}{2}\sigma^2\right]\tau}{\sigma\sqrt{\tau}}\right)$$

cannot be inverted in closed form in terms of P_t (with the exception of an at-the-money option). Instead, we will calculate numerical values for $\Psi(c, \sigma)$. We evaluate $\Psi(c_i, \sigma_j)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, u$ spanning (with the desired degree of accuracy) all possible values of c and σ, thus generating an n x u matrix of *implied prices*:

$$\Psi(c, \sigma) = \begin{bmatrix} \Psi(c_1, \sigma_1) & \Psi(c_1, \sigma_2) & \ldots & \Psi(c_1, \sigma_u) \\ \Psi(c_2, \sigma_1) & \Psi(c_2, \sigma_2) & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ \Psi(c_n, \sigma_1) & \ldots & \ldots & \Psi(c_n, \sigma_u) \end{bmatrix} \tag{6.18}$$

To ensure conformability in the calculations to follow, we also generate an n x u matrix for Σ of the form:

$$\Sigma = \begin{bmatrix} \sigma_1 & \ldots & \sigma_u \\ \ldots & \ldots & \ldots \\ \sigma_1 & \ldots & \sigma_u \end{bmatrix} \tag{6.19}$$

Substituting $\Psi(c, \sigma)$ and Σ in the formula for the joint density (i.e. equation (6.7)), we obtain an n x u matrix of values for $pdf(c, \sigma)$:

$$pdf(c, \sigma) = \begin{bmatrix} pdf(c_1, \sigma_1) & \ldots & pdf(c_1, \sigma_u) \\ \ldots & \ldots & \ldots \\ pdf(c_n, \sigma_1) & \ldots & pdf(c_n, \sigma_u) \end{bmatrix} \tag{6.20}$$

It should be noted that all the products between matrices that occur in the calculation of $pdf(c, \sigma)$ are Hadamard (elementwise) products.

It is now straightforward to obtain $pdf(c)$:

$$pdf(c) \approx \sum_{j=1}^{u} pdf(c_i, \sigma_j)\Delta j \tag{6.21}$$

Turning to the posterior density $pdf(c \backslash P_t, s)$, given in equation (6.14), we need to evaluate $\Theta(c)$. Remember, $\Theta(c)$ is the inverse of the option price c with respect to σ. Call it the

[16]For stylized facts in volatility see, for example, Ghysels *et al.* (1996).

implied volatility hereafter. However, the BS formula cannot be inverted in closed form in terms of σ (with the exception of an at-the-money option). Instead, we calculate numerical values for $\Theta(c)$. We evaluate $\Theta(c_i)$ for $i = 1, \ldots, n$ spanning with the desired degree of accuracy the range of values of c, thus generating an n-dimensional vector of implied volatilities.

6.4 Results

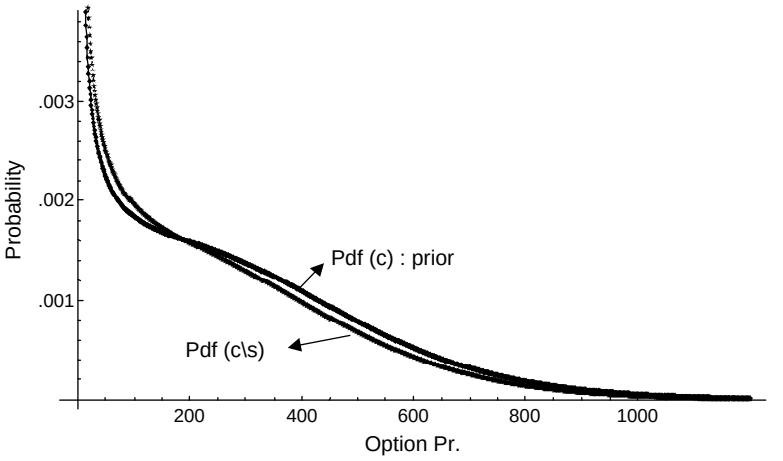
Typical values of the volatility of a stock index are in the range of 15% to 35% per annum. Assuming that time is measured in trading days and that there are 252 trading days per year, we have calculated (from prior 30-day data) that the expected rate of return of the FTSE 100 index is 15% per annum. We have also calculated that the volatility is 25% per annum. The standard error of our estimate is approximately $0.25/\sqrt{2 \times 30} = 3.2\%$ per annum. Thus at time 0, our 'prior' information (in daily format) is that: $E(\mu) = m = 0.0006$, $E(\sigma) = 0.0158$, Std. Dev.$(\sigma) = 0.002$, $\text{Var}(\sigma) = 4^{-06}$. Also using equation (6.21) we calculate $E(\sigma^2) = 2.5364^{-04}$. Since we know $E(\sigma)$ and $E(\sigma^2)$, we can now calculate values for the prior parameters λ, and θ using equations (6.20) and (6.19). To sum up, our 'prior beliefs' (in daily format) are: $\lambda = 0.004$, $\theta = 16.72$, and $m = 0.0006$. Also, the value of the index at time 0 is $P_0 = 2200$.

Turning to the sample information, we have that at time $t = 30$ the value of the index is $P_t = 2206$, the daily sample standard deviation is $s = 0.016$, and $\nu = t - 1 = 29$. Our data are chosen to conform with values presented in Ncube and Satchell (1997). Finally, the market information is as follows: consider a time t European Call option on the FTSE 100 index with exercise prices $K = 2025$, $K = 2225$ and $K = 2425$, and $\tau = 15$ (i.e. 15 trading days to maturity). The risk-free rate is $r = 0.0002$ (daily).

In Figure 6.1 we plot the prior density of the option price and the density of the option price conditional on the sample estimate of volatility for the case $K = 2025$. (Note that the latter density is for illustrative purposes, rather than of any practical use or theoretical significance.) Observe that *not conditioning* on the asset price induces a very large dispersion in the option price. This effect is magnified, since we are looking ahead 30 trading days ($t = 30$). Note also that conditioning on the sample estimate of volatility, when the underlying price is unknown, does not offer much improvement in reducing the dispersion of the option price. In Table 6.1 we report 2.5% and 97.5% quantiles for the densities exhibited in Figure 6.1. To illustrate the effect of how the dispersion of the option price decreases as the conditioning horizon decreases, we also report quantiles for the cases $t = 25$, 20, 15, 10 and 5. A graphical illustration of the prior density of the BS option price for varying values of t is exhibited in Figure 6.2.

At the bottom of Table 6.1 we also report summary statistics for the distribution of the underlying (i.e. the log-normal) distribution for $P_0 = 2200$, $t = 30$ and for $\mu = m = 0.0006$ and $\sigma = s = 0.016$. To this end, we report the values of the option price that correspond to the 2.5% and 97.5% quantiles of the distribution of the asset price. Note that in order to calculate BS prices, we assume that volatility is known and equal to its sample estimate (i.e. $\sigma = s = 0.016$).

From Table 6.1, it is interesting to observe that (once we condition on s) the true 95% range of the option price, given by $pdf(c \backslash s)$, is wider than the one we would obtain if we took advantage of the monotonicity properties of the BS option price with respect to the



$c = C(P_t, \sigma)$: P_t : unknown, σ : unknown

$t = 30$, $P_0 = 2200$, $K = 2025$, $\tau = 15$, $r = 0.0002$, $\lambda = 0.004$, $\theta = 16.72$, $m = 0.0006$, $s = 0.016$, $v = 29$

Figure 6.1 Prior and conditional on the sample estimate of volatility probability density functions of the BS option price.

Table 6.1 2.5% and 97.5% quantiles of $pdf(c)$ and $pdf(c\backslash s)$ for $t = 30$, 25, 20, 15, 10 and 5

	Quantiles		
$pdf(c)$:**Prior**	**0.025**	**0.975**	**Mean of $pdf(c)$**
$t = 30$	0.1	828.9	262.9
$t = 25$	0.4	751.7	252.2
$t = 20$	1.7	679.1	240.4
$t = 15$	5.2	606.5	227.6
$t = 10$	14.9	523.4	214.1
$t = 5$	39.8	416.2	200.2
	Quantiles		
$pdf(c\backslash s)$	**0.025**	**0.975**	**Mean of $pdf(c\backslash s)$**
$t = 30$	0.1	786.9	234.1
$t = 25$	0.1	712.0	227.6
$t = 20$	0.8	646.2	220.3
$t = 15$	3.6	583.0	212.4
$t = 10$	11.5	509.1	204.1
$t = 5$	37.2	410.4	195.6

Summary statistics for the *log-normal* distribution ($P_0 = 2200$, $t = 30$, $\mu = 0.0006$, $\sigma = 0.016$) and BS prices corresponding to the 2.5% and 97.5% quantiles of the log-normal distribution ($P_t = 1879.2$ and $P_t = 2649.5$, $K = 2025$, $\tau = 15$, $\sigma = 0.016$, $r = 0.0002$)

(Continued)

Table 6.1 Continued

	Quantiles					
	0.025	**0.975**	**Mean**	**SD**	**Skewness**	**Exs kurtosis**
$pdf(P_t \backslash P_0, \mu, \sigma, t)$	1879.2	2649.5	2240	196.7	0.264	0.124
BS price	7.38	630.54				

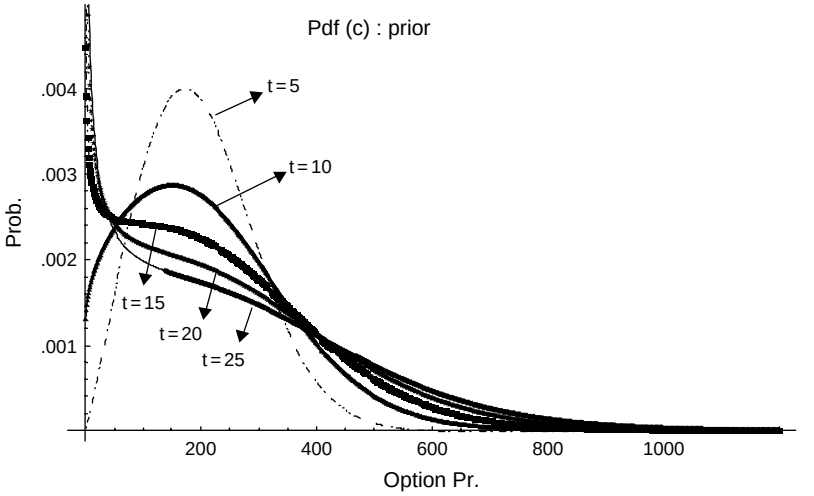
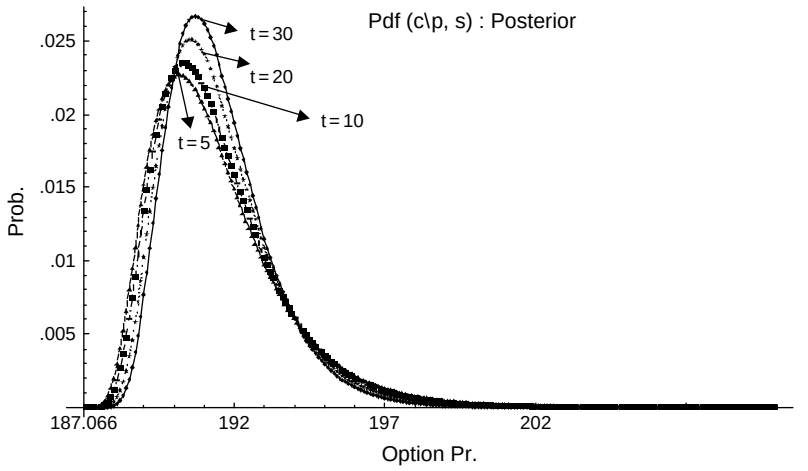


Figure 6.2 Prior densities for the BS option price for varying values of t.

underlying, and used the log-normal distribution to derive 95% confidence intervals for the option price. (Compare a range of (0.1, 786.9) with (7.38, 630.54).) To do the latter, one has to assume that randomness arises from the asset price while volatility is known and equal to its sample estimate. This is the approach of Ncube and Satchell (1997). If we do not condition on s, the true 95% range of the option price, given by *pdf(c)*, is even wider (i.e. (0.1, 828.9)).

We now turn to the posterior density where we condition on the asset price P_t and on the sample estimate of volatility s. Let us first illustrate the effect of varying t values for the posterior density. In Figure 6.3 we plot the posterior density of the BS option price for $t = 30$, 20, 10 and 5, and $K = 2025$, $P_t = 2206$ and $s = 0.016$ (everything else as already defined above). This time we observe the opposite effect of what we saw for the prior density – that is, the larger the sample size t the smaller the dispersion in the option price (see Table 6.2). Indeed, a large sample size will provide a better estimate for the volatility (provided that the sample size is not too large, to avoid issues of non-stationarity) and hence reduce estimation risk. Despite the fact that for a large t the prior density will



$c = C(P_t, \sigma) : P_t = 2206, \sigma$: unknown

$P_0 = 2200, K = 2025, \tau = 15, r = 0.0002, s = 0.016, v = 29, \lambda = 0.004, \theta = 16.72, m = 0.0006.$

Figure 6.3 Posterior densities of the BS option price for varying values of t.

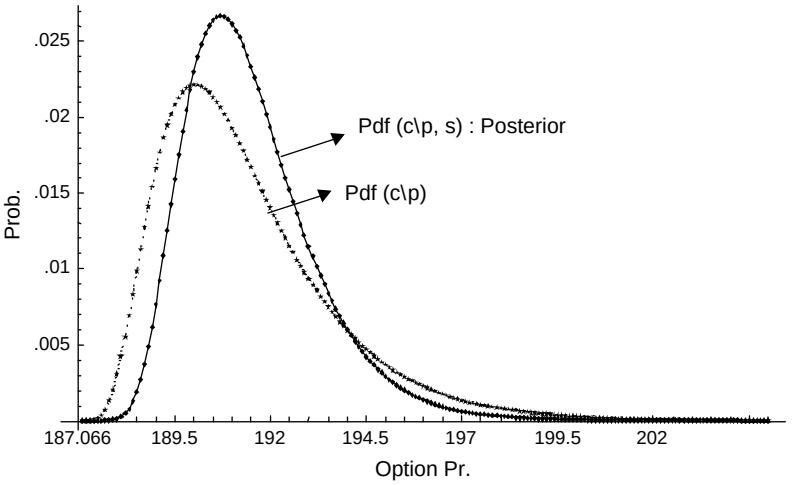
Table 6.2 2.5% and 97.5% quantiles of $pdf(c\backslash P_t, s)$ for $t = 30$, 20, 10 and 5

	Quantiles		
	0.025	**0.975**	**Mean* of $pdf(c\backslash P_t, s)$**
$t = 30$	188.9	195.2	189.8
$t = 20$	188.7	196.4	189.8
$t = 10$	188.5	196.9	190.1
$t = 5$	188.3	197.2	190.4

*If we combine the posterior density with a quadratic loss function, the mean of the posterior distribution is an optimal point estimate.

be less informative (see Figure 6.2), the sample information is more robust and this is reflected in the posterior density.

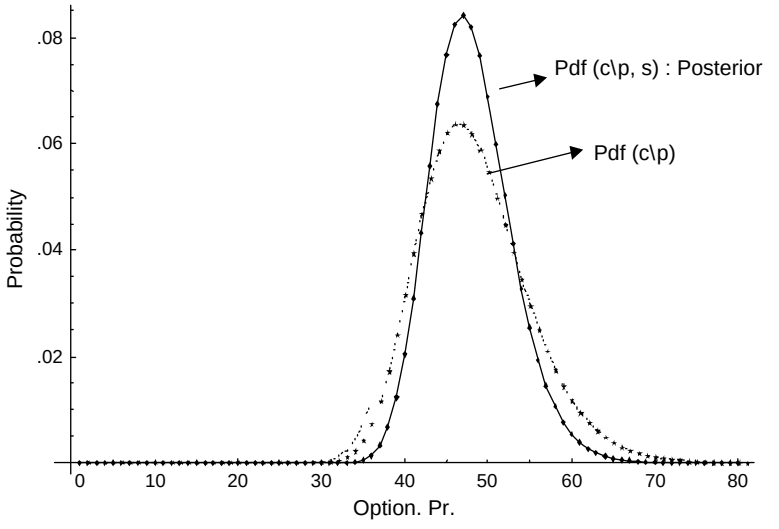
Comparing Figures 6.1 and 6.2 with Figure 6.3, it is obvious that conditioning on the asset price dramatically reduces the variability of the option price. We now present graphs to illustrate how the dispersion of the option price changes from conditioning on the asset price only, to conditioning on both the asset price and the sample estimate of volatility. In other words, we compare the density of the option price conditional on the asset price (i.e. $pdf(c\backslash P_t)$) with the posterior density (i.e. $pdf(c\backslash P_t, s)$). We do this for an in-the-money option (i.e. $K = 2025$ and $P_t = 2206$), a near-the-money option (i.e. $K = 2225$ and $P_t = 2206$), an at-the-money option (i.e. $K = 2212.63$ and $P_t = 2206$), and an out-of-the-money option (i.e. $K = 2425$ and $P_t = 2206$). We present our results, complete with summary statistics for each distribution, in Figures 6.4, 6.5 and 6.6, and Tables 6.3, 6.4 and 6.5 respectively.



$c = C(P_t, \sigma) : P_t = 2206, \sigma$: unknown

$P_0 = 2200, K = 2025, \tau = 15, r = 0.0002, t = 30, s = 0.016, v = 29, \lambda = 0.004, \theta = 16.72, m = 0.0006.$

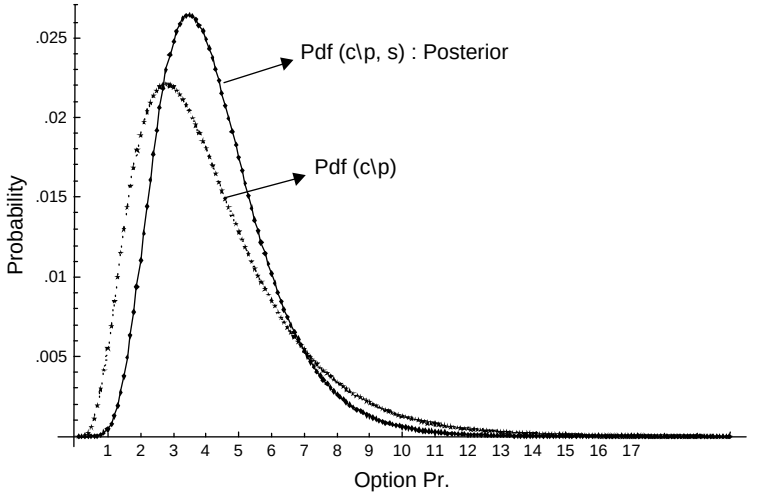
Figure 6.4 Posterior and conditional (on the asset price) probability density functions for the BS option price: in-the-money case.



$c = C(P_t, \sigma) : P_t = 2206, \sigma$: unknown

$P_0 = 2200, K = 2225, \tau = 15, r = 0.0002, t = 30, s = 0.016, v = 29, \lambda = 0.004, \theta = 16.72, m = 0.0006.$

Figure 6.5 Posterior and conditional (on the asset price) probability density functions for the BS option price: near-the-money case.



$c = C(P_t, \sigma) : P_t = 2206, \sigma$: unknown

$P_0 = 2200, K = 2425, \tau = 15, r = 0.0002, t = 30, s = 0.016, v = 29, \lambda = 0.004, \theta = 16.72, m = 0.0006.$

Figure 6.6 Posterior and conditional (on the asset price) probability density functions for the BS option price: out-of-the-money case.

Table 6.3 Summary statistics of the distributions exhibited in Figure 6.4

	Quantiles					
	0.025	**0.975**	**Mean**	**SD**	**Skewness**	**Excess kurtosis**
$pdf(c\backslash P_t, s)$	188.9	195.2	189.8	2.3	1.81	1.12
$pdf(c\backslash P_t)$	188.3	197.2	191.2	2.3	1.52	3.00

Table 6.4 Summary statistics of distributions exhibited in Figure 6.5 and at-the-money case

	Quantiles					
	0.025	**0.975**	**Mean**	**SD**	**Skewness**	**Excess kurtosis**
			Near-the-money			
$pdf(c\backslash P_t, s)$	39.0	59.1	47.6	4.85	0.70	0.57
$pdf(c\backslash P_t)$	35.8	61.6	47.3	6.60	0.62	0.64
At-the-money: $P_t = K\exp(-r\tau)$. ($K = 2212.63$, **everything else as in Figure 6.5**)						
$pdf(c\backslash P_t, s)$	44.7	64.9	53.3	4.86	0.74	0.61
$pdf(c\backslash P_t)$	41.4	73.9	53.1	6.62	0.64	0.78

Table 6.5 Summary statistics of distributions exhibited in Figure 6.6

	Quantiles					
	0.025	0.975	Mean	SD	Skewness	Excess kurtosis
$pdf(c\backslash P_t, s)$	1.7	8.9	4.27	1.73	1.15	2.14
$pdf(c\backslash P_t)$	1.1	10.1	4.22	2.30	1.45	3.24

Note that $pdf(c\backslash P_t, s)$ and $pdf(c\backslash P_t)$ are defined within the support of the distribution. For example, for the case $K = 2025$ the density has the support given by the no-arbitrage bounds of the option price: $187.066 < c = C(P_t, \sigma) < 2206$. For the cases $K = 2225$ and $K = 2425$, the support is $0 < c = C(P_t, \sigma) < 2206$.

Finally, and perhaps most importantly, note that the posterior distributions for all three cases generally exhibit excess kurtosis and are positively skewed when compared with the normal distribution. In particular for the at-the-money and near-the-money cases the distributions are close to normal, however as we move progressively out- or in- the money the posterior distribution of the option price exhibits an increasingly thinner left tail than the normal distribution (see Tables 6.3–6.5).

6.5 Concluding remarks and issues for further research

European option prices depend on five factors, mamely the underlying price, exercise price, volatility, risk-free rate and time to maturity. There is no randomness arising in the exercise price, time to maturity and risk-free rate (if we do not consider the stochastic interest rate case). However, the randomness arising from either price uncertainty or estimation error of volatility is of great interest, as we do not know the underlying price in the future and true volatility, and they must be estimated from historical data or via other methods.

In the foregoing Bayesian analysis we have derived the true distribution of the Black–Scholes option price with the randomness arising from both the underlying asset price and its volatility. To this end we have illustrated (see Figures 6.4–6.6 and Tables 6.3–6.5) that the (posterior) distribution for a long call option exhibits positive skewness and excess kurtosis with the departure from normality becoming more profound as the option moves progressively from at-the-money to either side away-from-the-money.

Regarding the behaviour of the distribution of the option price as a function of the information flow, the results presented in Figures 6.1–6.6 and Tables 6.1–6.5 show the extent to which conditioning on the asset price dramatically reduces the variability of the option price. Indeed, since as a Bayesian problem the BS option price depends on both parameters (volatility:(σ)) and data (price: (P_t)), not conditioning on the data induces a very large dispersion in the option price. It should, however, be mentioned that as the conditioning horizon decreases (i.e. the time between the initial price P_0 and the final price P_t), variability in the option price gradually decreases as well. If we include the sample variance (s^2) into the data, not conditioning on it does not have as dramatic an impact as not conditioning on the asset price, but we show that conditioning on both results in less variability for the option price than conditioning only on price.

This will have important implications for forecasting. Although this chapter is not about forecasting, we see this analysis as a necessary prelude to establishing a Bayesian theory of option price forecasting. Existing theories (such as those of Karolyi, 1993; Noh *et al.*, 1994; Hwang and Satchell, 1998; etc.) use only the implied volatility (or other measures of volatility (e.g. GARCH)) to forecast option prices while keeping the price of the underlying fixed. Our theory will allow us to consider forecasting when both prices and volatility can vary, as they do in many practical applications. For a forecasting application/extension of the underlying theory developed in this chapter, we refer the reader to Darsinos and Satchell (2001a). There we utilize the Bayesian approach to combine implied and historical volatility information and forecast the prices of FTSE 100 Index European options.

Furthermore, our results have potential uses in risk management as we can report VaR (Value at Risk) and other distributional measures. Having derived in analytic form the 'true' distribution of the Black–Scholes option price, we have provided a viable and potentially superior alternative to the approaches that use linear (delta normal) or quadratic (delta-gamma) approximations for the calculation of VaR. We illustrated that the true distribution of a long call option tends to have an increasingly thinner left tail than the normal distribution (similarly, the distribution of a short call will tend to have an increasingly fatter tail). The VaR for an asset or portfolio of assets is critically dependent on the left tail of their distributions. If, for example, one assumes that the distribution of a long call option is normal, then the tendency will be to calculate a VaR that is higher than the true VaR. Similarly, for a short call the calculated VaR will be too low. Although we do not consider portfolio problems, it is possible to carry out such extensions. For an application/extension of our theory in value-at-risk (VaR) calculations and a comparison with existing VaR approaches, see Darsinos and Satchell (2001b).

Likewise, one could use our methodology in option pricing models other than the Black–Scholes. Thus, at least in principle, we could incorporate randomness due to interest rates (Merton, 1973) or specific models of volatility (such as Duan, 1995; Bauwens and Lubrano, 2000). Finally, there is a class of financial problems involving the valuation of warrants and corporate bonds. These problems require the determination of the distribution of the return of an asset, which is the combination of a value process and an option on that value process. Our analysis will allow us to address such questions, and we refer the reader to Darsinos and Satchell (2001c), where we derive the distribution of the stock price for a firm that issues warrants and/or executive stock options and thus provide an alternative approach for warrant valuation.

Acknowledgements

Theofanis Darsinos gratefully acknowledges financial support from the A.G. Leventis Foundation, the Wrenbury Scholarship Fund at the University of Cambridge, and the Economic and Social Research Council (ESRC).

Appendix

A.1

Proposition 1: The joint unconditional density of the price P_t and volatility σ is given by:

$$pdf(P_t, \sigma) = \frac{1}{\sqrt{\pi t} P_t} \frac{\lambda^\theta}{\Gamma(\theta)} \frac{1}{\sigma^{2(\theta+1)}} \exp\left(-\frac{\lambda}{\sigma^2}\right) \exp\left(-\frac{1}{4\sigma^2 t}\left[\ln\left(\frac{P_t}{P_0}\right) - \left(m - \frac{1}{2}\sigma^2\right)t\right]^2\right)$$

Proof of Proposition:

From Assumptions 4 and 2 we have expressions for $pdf(\sigma)$ and $pdf(\mu \backslash \sigma)$ respectively. Then $pdf(\mu, \sigma)$ is just $pdf(\sigma)pdf(\mu \backslash \sigma)$:

$$pdf(\mu, \sigma) = \sqrt{\frac{2t}{\pi}} \frac{\lambda^\theta}{\Gamma(\theta)} \frac{1}{\sigma^{2(\theta+1)}} \exp\left(-\frac{t(\mu - m)^2}{2\sigma^2}\right) \exp\left(-\frac{\lambda}{\sigma^2}\right)$$

Similarly $pdf(P_t, \mu, \sigma)$ is given by $pdf(\mu, \sigma)pdf(P_t \backslash \mu, \sigma)$.

We have just obtained $pdf(\mu, \sigma)$, and in Assumption 1 we state $pdf(P_t \backslash \mu, \sigma)$. Then

$$pdf(P_t, \mu, \sigma) = \frac{1}{\pi P_t} \frac{\lambda^\theta}{\Gamma(\theta)} \frac{1}{\sigma^{2\theta+3}} \times \exp\left(-\frac{\lambda}{\sigma^2}\right) \exp\left\{-\frac{1}{2\sigma^2 t}\left(\left[\ln\left(\frac{P_t}{P_0}\right) + \frac{\sigma^2 t}{2} - \mu t\right]^2 + t^2(\mu - m)^2\right)\right\}$$

To get the proposed result we therefore need to integrate out μ:

$$\begin{aligned} pdf(P_t, \sigma) &= \int_{-\infty}^{\infty} pdf(P_t, \mu, \sigma) \partial\mu \\ &= K \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2 t}\left((z - \mu t)^2 + t^2(\mu - m)^2\right)\right\} \partial\mu \end{aligned}$$

where we have set $K = \frac{1}{\pi P_t} \frac{\lambda^\theta}{\Gamma(\theta)} \frac{1}{\sigma^{2\theta+3}} \exp\left(-\frac{\lambda}{\sigma^2}\right)$ and $z = \ln\left(\frac{P_t}{P_0}\right) + \frac{1}{2}\sigma^2 t$.

Let us now evaluate the integral:

$$\begin{aligned} &\int_{-\infty}^{\infty} \exp\left\{-\frac{(z - \mu t)^2 + t^2(\mu - m)^2}{2\sigma^2 t}\right\} \partial\mu \\ &= \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2 t}\left[(z^2 + t^2 m^2) + \mu^2(2t^2) - 2\mu t(z + mt)\right]\right) \partial\mu \\ &= \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2 t}\left[z^2 + t^2 m^2 + 2t^2\left\{\mu^2 - 2\mu\left(\frac{z + mt}{2t}\right) + \left(\frac{z + mt}{2t}\right)^2\right\}\right.\right. \end{aligned}$$

$$
\begin{aligned}
&-2t^2\left(\frac{z+mt}{2t}\right)^2\Bigg]\Bigg)\partial\mu \\
&=\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2\sigma^2 t}\left[\frac{1}{2}(z-tm)^2+2t^2\left(\mu-\frac{z+mt}{2t}\right)^2\right]\right)\partial\mu \\
&=\exp\left(-\frac{1}{4\sigma^2 t}(z-tm)^2\right)\int_{-\infty}^{\infty}\exp\left(-\frac{t}{\sigma^2}\left(\mu-\frac{z+mt}{2t}\right)^2\right)\partial\mu \\
&=\exp\left(-\frac{1}{4\sigma^2 t}(z-tm)^2\right)\sqrt{2\pi}\left(\frac{\sigma}{\sqrt{2t}}\right) \\
&\quad\times\left[\frac{1}{\sqrt{2\pi}\left(\frac{\sigma}{\sqrt{2t}}\right)}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2(\sigma^2/2t)}\left(\mu-\frac{z+mt}{2t}\right)^2\right)\partial\mu\right] \\
&=\exp\left(-\frac{1}{4\sigma^2 t}(z-tm)^2\right)\left(\sqrt{\frac{\pi}{t}}\right)\sigma
\end{aligned}
$$

Therefore:

$$
pdf(P_t,\sigma)=K\exp\left(-\frac{1}{4\sigma^2 t}(z-tm)^2\right)\left(\sqrt{\frac{\pi}{t}}\right)\sigma
$$

Substituting in the values for K and z, we get the proposed result for $pdf(P_t,\sigma)$.

A.2

Proposition 3: The unconditional density of the statistic s is given by:

$$
pdf(s)=\frac{2}{\mathrm{B}\left(\frac{\nu}{2},\theta\right)}\lambda^{\theta}\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}\frac{s^{\nu-1}}{\left(\lambda+\frac{\nu s^2}{2}\right)^{\frac{\nu}{2}+\theta}}
$$

Proof of Proposition:

We start by obtaining $pdf(s^2,\sigma^2)=pdf(\sigma^2)pdf(s^2\backslash\sigma^2)$, where $pdf(\sigma^2)$ is given in Assumption 4 and $pdf(s^2\backslash\sigma^2)$ in Assumption 3. Then:

$$
pdf(s^2,\sigma^2)=\frac{\lambda^{\theta}}{\Gamma(\theta)}\frac{\nu^{\frac{\nu}{2}}(s^2)^{\frac{\nu}{2}-1}}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)}\left(\frac{1}{\sigma^2}\right)^{\frac{\nu}{2}+\theta+1}\exp\left(-\frac{2\lambda+\nu s^2}{2\sigma^2}\right)
$$

Now let

$$
K=\frac{\lambda^{\theta}}{\Gamma(\theta)}\frac{\nu^{\frac{\nu}{2}}(s^2)^{\frac{\nu}{2}-1}}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)},\ m=\frac{\nu}{2}+\theta+1,\ a=\frac{2\lambda+\nu s^2}{2}.
$$

Then

$$pdf(s^2) = \int_0^\infty pdf(s^2, \sigma^2)\partial\sigma^2 = K\int_0^\infty \left(\frac{1}{\sigma^2}\right)^m \exp\left(-\frac{a}{\sigma^2}\right)\partial\sigma^2$$

We need to evaluate:

$$\int_0^\infty \left(\frac{1}{\sigma^2}\right)^m \exp\left(-\frac{a}{\sigma^2}\right)\partial\sigma^2$$

Let $\frac{1}{\sigma^2} = x \Rightarrow -(\frac{1}{\sigma^2})^2\partial\sigma^2 = \partial x$. Then

$$\int_0^\infty \left(\frac{1}{\sigma^2}\right)^m \exp\left(-\frac{a}{\sigma^2}\right)\partial\sigma^2 = -\int_\infty^0 x^{m-2}\exp(-ax)\partial x = \int_0^\infty x^{m-2}\exp(-ax)\partial x.$$

Now

$$\int_0^\infty x^{m-2}\exp(-ax)\partial x = \left[-\frac{1}{a}x^{m-2}\exp(-ax)\right]_0^\infty - \frac{m-2}{-a}\int_0^\infty x^{m-3}\exp(-ax)\partial x =$$

note that $\lim_{x\to\infty}(x^k e^{-x}) = 0$

$$= \frac{m-2}{a}\int_0^\infty x^{m-3}\exp(-ax)\partial x$$

$$= \frac{m-2}{a}\left\{\left[-\frac{1}{a}x^{m-3}\exp(-ax)\right]_0^\infty - \frac{m-3}{-a}\int_0^\infty x^{m-4}\exp(-ax)\partial x\right\}$$

$$= \frac{(m-2)(m-3)}{a^2}\int_0^\infty x^{m-4}\exp(-ax)\partial x = \ldots\ldots\ldots = \frac{(m-2)!}{a^{m-2}}\int_0^\infty \exp(-ax)\partial x$$

$$= -\frac{(m-2)!}{a^{m-1}}[\exp(-ax)]_0^\infty = \frac{(m-2)!}{a^{m-1}}$$

$$= \frac{\left(\frac{\upsilon}{2}+\theta-1\right)!}{\left(\frac{2\lambda+\upsilon s^2}{2}\right)^{\frac{\upsilon}{2}+\theta}}$$

$$\Rightarrow pdf(s^2) = K\frac{\Gamma\left(\frac{\upsilon}{2}+\theta\right)}{\left(\lambda+\frac{\upsilon s^2}{2}\right)^{\frac{\upsilon}{2}+\theta}}$$

Substituting for K we get:

$$pdf(s^2) = \frac{\lambda^\theta}{\Gamma(\theta)}\frac{\upsilon^{\frac{\upsilon}{2}}(s^2)^{\frac{\upsilon}{2}-1}}{2^{\frac{\upsilon}{2}}\Gamma\left(\frac{\upsilon}{2}\right)}\frac{\Gamma\left(\frac{\upsilon}{2}+\theta\right)}{\left(\lambda+\frac{\upsilon s^2}{2}\right)^{\frac{\upsilon}{2}+\theta}}$$

$$= \frac{1}{\mathrm{B}\left(\frac{\upsilon}{2},\theta\right)}\lambda^\theta\left(\frac{\upsilon}{2}\right)^{\frac{\upsilon}{2}}\frac{(s^2)^{\frac{\upsilon}{2}-1}}{\left(\lambda+\frac{\upsilon s^2}{2}\right)^{\frac{\upsilon}{2}+\theta}}.$$ [18]

[18]This implies that $\frac{\upsilon s^2}{2}$ is unconditionally distributed inverted-beta $f_{i\beta 2}(\frac{\upsilon s^2}{2}\backslash\frac{\upsilon}{2},\theta,\lambda) \equiv \frac{1}{\mathrm{B}(\frac{\upsilon}{2},\theta)}\lambda^\theta\frac{(\frac{\upsilon s^2}{2})^{\frac{\upsilon}{2}-1}}{(\lambda+\frac{\upsilon s^2}{2})^{\frac{\upsilon}{2}+\theta}}$.

Having obtained $pdf(s^2)$, it is straightforward to get $pdf(s)$:

$$pdf(s) = 2spdf(s^2) = \frac{2}{\mathrm{B}\left(\frac{\upsilon}{2}, \theta\right)} \lambda^{\theta} \left(\frac{\upsilon}{2}\right)^{\frac{\upsilon}{2}} \frac{s^{\upsilon-1}}{\left(\lambda + \frac{\upsilon s^2}{2}\right)^{\frac{\upsilon}{2}+\theta}}$$

A.3

Proposition 4: The conditional (on the sample and prior information) density of the underlying asset price is given by:

$$pdf(P_t \backslash s) = \frac{K_{\theta+\frac{t}{2}}\left(\frac{1}{4}\sqrt{2(2\lambda+\upsilon s^2)t + (\ln(P_t/P_0) - mt)^2}\right)}{\sqrt{\pi t} P_t \Gamma\left(\theta + \frac{\upsilon}{2}\right)}$$
$$\times \left(\lambda + \frac{\upsilon}{2}s^2\right)^{\theta+\frac{\upsilon}{2}} \left(\frac{t}{16}\right)^{\theta+\frac{t}{2}} \left[\frac{2(2\lambda+\upsilon s^2)t + (\ln(P_t/P_0) - mt)^2}{64}\right]^{\frac{2\theta+t}{4}}$$
$$\times \exp\left(-\frac{\ln(P_t/P_0) - mt}{4}\right)$$

where $K_n(\,)$ is the modified Bessel function of the second kind of order n.

Proof of Proposition:

Let us first illustrate how to derive the marginal density of the asset price (i.e. $pdf(P_t)$). Then, following the same procedure it is straightforward to obtain the conditional density of the asset price (i.e. $pdf(P_t \backslash s)$), we have that $pdf(P_t) = \int_0^{\infty} pdf(P_t, \sigma^2)\partial\sigma^2$

where

$$pdf(P_t, \sigma^2) = \frac{1}{2\sqrt{\pi t} P_t} \frac{\lambda^{\theta}}{\Gamma(\theta)} \left(\frac{1}{\sigma^2}\right)^{\theta+\frac{3}{2}} \exp\left(-\frac{\lambda}{\sigma^2}\right) \exp\left(-\frac{1}{4\sigma^2 t}\left[\ln\left(\frac{P_t}{P_0}\right) - \left(m - \frac{1}{2}\sigma^2\right)t\right]^2\right)$$

(this follows straightforwardly from Proposition 1)

Now let $A = \frac{1}{2\sqrt{\pi t} P_t} \frac{\lambda^{\theta}}{\Gamma(\theta)}$ and $z = \ln\left(\frac{P_t}{P_0}\right) - mt$. Then

$$pdf(P_t, \sigma^2) = A\left(\frac{1}{\sigma^2}\right)^{\left(\theta+\frac{1}{2}\right)+1} \exp\left(-\frac{\lambda}{\sigma^2}\right) \exp\left(-\frac{1}{4\sigma^2 t}\left[z - \frac{1}{2}\sigma^2 t\right]^2\right)$$
$$= A\left(\frac{1}{\sigma^2}\right)^{\left(\theta+\frac{1}{2}\right)+1} \exp\left(-\frac{\lambda}{\sigma^2}\right) \exp\left(-\frac{z^2}{4\sigma^2 t} - \frac{z}{4} - \frac{\sigma^2 t}{16}\right)$$

$$= A\exp\left(-\frac{z}{4}\right)\left(\frac{1}{\sigma^2}\right)^{\left(\theta+\frac{1}{2}\right)+1}\exp\left[-\frac{\left(\lambda+z^2/4t\right)}{\sigma^2}\right]\exp\left(-\frac{t}{16}\sigma^2\right)$$

Also let $c = \lambda + \frac{z^2}{4t}$ and $p = \frac{t}{16}$. Then

$$pdf(P_t) = \int_0^\infty pdf(P_t, \sigma^2)\partial\sigma^2$$

$$= A\exp\left(-\frac{z}{4}\right)\frac{\Gamma\left(\theta+\frac{1}{2}\right)}{c^{\theta+\frac{1}{2}}}\int_0^\infty \underbrace{\frac{c^{\theta+\frac{1}{2}}}{\Gamma\left(\theta+\frac{1}{2}\right)}\left(\frac{1}{\sigma^2}\right)^{\left(\theta+\frac{1}{2}\right)+1}\exp\left[-\frac{c}{\sigma^2}\right]}_{Inverted-Gamma-1\ \text{function}}\exp(-p\sigma^2)\partial\sigma^2$$

Observe now that the required integral is the Laplace transform of an inverted-gamma function:

$$L\left(f_{i\gamma}\right) = F(p) = \int_0^\infty f_{i\gamma}(\sigma^2)e^{-p\sigma^2}d\sigma^2$$

Omitting some tedious algebra to calculate the Laplace transform, we arrive at the result:

$$pdf(P_t) = A\exp\left(-\frac{z}{4}\right)2p^{\theta+1/2}(cp)^{-\frac{2\theta+1}{4}}K_{\theta+\frac{1}{2}}(2\sqrt{cp})$$

Substituting in the values for A, z, p and c we get the marginal density of the asset price:

$$pdf(P_t) = \frac{K_{\theta+\frac{1}{2}}\left(\frac{1}{4}\sqrt{4\lambda t + (\ln(P_t/P_0) - mt)^2}\right)}{\sqrt{\pi t}P_t\Gamma(\theta)}\lambda^\theta\left(\frac{t}{16}\right)^{\theta+\frac{1}{2}}\left[\frac{4\lambda t + (\ln(P_t/P_0) - mt)^2}{64}\right]^{\frac{2\theta+1}{4}}$$

$$\times\exp\left(-\frac{\ln(P_t/P_0) - mt}{4}\right)$$

where $K_{\theta+\frac{1}{2}}(\frac{1}{4}\sqrt{4\lambda t + (\ln(P_t/P_0) - mt)^2})$ is the modified Bessel function of the second kind of order $\theta + \frac{1}{2}$.

Now to derive $pdf(P_t \backslash s)$ we need to follow a similar procedure to the one outlined above. This time we calculate

$$pdf(P_t \backslash s) = \int_0^\infty pdf(P_t, \sigma^2 \backslash s)\partial\sigma^2$$

Note that $pdf(P_t, \sigma^2 \backslash s) = \frac{pdf(P_t, \sigma^2, s)}{pdf(s)}$, where the numerator can be straightforwardly obtained from Proposition 2 and the denominator is given in Proposition 3. Again omitting the algebra, the result is:

$$pdf(P_t \backslash s) = \frac{K_{\theta+\frac{t}{2}}\left(\frac{1}{4}\sqrt{2(2\lambda+\upsilon s^2)t+(\ln(P_t/P_0)-mt)^2}\right)}{\sqrt{\pi t}P_t\Gamma\left(\theta+\frac{\upsilon}{2}\right)} \times \left(\lambda+\frac{\upsilon}{2}s^2\right)^{\theta+\frac{\upsilon}{2}}\left(\frac{t}{16}\right)^{\theta+\frac{t}{2}}\left[\frac{2(2\lambda+\upsilon s^2)t+(\ln(P_t/P_0)-mt)^2}{64}\right]^{\frac{2\theta+t}{4}} \times \exp\left(-\frac{\ln(P_t/P_0)-mt}{4}\right)$$

A.4

1. We want to derive the density of the option price conditional on the sample estimate of volatility: i.e. $pdf(c \backslash s)$.

   Consider first $pdf(P_t, \sigma \backslash s) = \dfrac{pdf(P_t, \sigma, s)}{pdf(s)}$.

   $pdf(P_t, \sigma, s)$ is given in Proposition 2 and $pdf(s)$ in Proposition 3. Hence

$$pdf(P_t, \sigma \backslash s) = \frac{1}{\sqrt{\pi t}P_t\Gamma\left(\frac{\upsilon}{2}+\theta\right)}\frac{\left(\lambda+\frac{\upsilon s^2}{2}\right)^{\frac{\upsilon}{2}+\theta}}{\sigma^{2\theta+\upsilon+2}}\exp\left(-\frac{2\lambda+\upsilon s^2}{2\sigma^2}\right) \times \exp\left(-\frac{1}{4\sigma^2 t}\left[\ln\left(\frac{P_t}{P_0}\right)-\left(m-\frac{1}{2}\sigma^2\right)t\right]^2\right)$$

   Using now the same transformation as we did for the prior density:

$$c = C(P_t, \sigma) \Rightarrow \Psi(c) = \Psi(c, \sigma) = P_t$$
$$\sigma = \sigma$$

   we get

$$pdf(c, \sigma \backslash s) = \frac{1}{\Phi(d_1^*)\sqrt{\pi t}\Psi(c, \sigma)\Gamma\left(\frac{\upsilon}{2}+\theta\right)}\frac{\left(\lambda+\frac{\upsilon s^2}{2}\right)^{\frac{\upsilon}{2}+\theta}}{\sigma^{2\theta+\upsilon+2}}\exp\left(-\frac{2\lambda+\upsilon s^2}{2\sigma^2}\right) \times \exp\left(-\frac{1}{4\sigma^2 t}\left[\ln\left(\frac{\Psi(c, \sigma)}{P_0}\right)-\left(m-\frac{1}{2}\sigma^2\right)t\right]^2\right)$$

where $d_1^* = \dfrac{\left[\ln\left(\dfrac{\Psi(c,\sigma)}{Ke^{-r\tau}}\right) + \dfrac{\sigma^2\tau}{2}\right]}{\sigma\sqrt{\tau}}$

Integrating out σ numerically will give us $pdf(c\backslash s)$.

2. We also want to derive the density of the option price conditional on the price: i.e. $pdf(c\backslash P_t)$.

   Consider first $pdf(\sigma\backslash P_t) = \dfrac{pdf(P_t,\sigma)}{pdf(P_t)}$.

   $pdf(P_t,\sigma)$ is given in Proposition 1 and $pdf(P_t)$ in Section A.3 above. Applying the transformation:

$$c = C(\sigma) \Leftrightarrow \sigma = \Theta(c)$$

we get:

$$pdf(c\backslash P_t) = \frac{vega}{K_{\theta+\frac{1}{2}}\left(\theta+\frac{1}{2}, \frac{1}{4}\sqrt{4\lambda t + \left(\ln\left(\frac{P_t}{P_0}\right) - mt\right)^2}\right)}$$

$$\times \frac{\exp\left(-\frac{\lambda}{\Theta[c]^2} - \frac{1}{4\Theta[c]^2 t}\left[\ln\left(\frac{P_t}{P_0}\right) - \left(m - \frac{1}{2}\Theta[c]^2\right)t\right]^2 + \left[\ln\left(\frac{P_t}{P_0}\right) - mt\right] \Big/ 4\right)}{\Theta[c]^{2(\theta+1)}\left(\frac{t}{16}\right)^{\theta+\frac{1}{2}}\left[\frac{4\lambda t + \left(\ln\left(\frac{P_t}{P_0}\right) - mt\right)^2}{64}\right]^{-\frac{2\theta+1}{4}}}$$

References

Bauwens, L. and Lubrano, M. (2000). Bayesian option pricing using asymmetric GARCH models. Discussion Paper, CORE, Université catholique de Louvain.

Black, F. (1976). Studies of stock price volatilities. *Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economics Section*, 56:177–181.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81:637–659.

Bollerslev, T. (1986). Generalised autoregressive conditional heteroscedasticity. *Journal of Econometrics*, 31:307–327.

Boyle, P. and Ananthanarayanan, A. (1977). The impact of variance estimation in option valuation models. *Journal of Financial Economics*, 5:375–387.

Butler, J. and Schachter, B. (1986). Unbiased estimation of the Black/Scholes formula. *Journal of Financial Economics*, 15:341–357.

Chiras, D.P. and Manaster, S. (1978). The information content of option prices and a test of market efficiency. *Journal of Financial Economics*, 6:213–234.

Christie, A. (1982). The stochastic behaviour of common stock variances. *Journal of Financial Economics*, 10:407–432.

Darsinos, T. and Satchell, S. (2001a). Bayesian forecasting of options prices. A natural framework for pooling historical and implied volatility information. DAE (Department of Applied Economics) Working Paper, University of Cambridge.
Darsinos, T. and Satchell, S. (2001b). A Bayesian approach to value at risk for derivatives. Mimeo.
Darsinos, T. and Satchell, S. (2001c). The implied distribution for stocks of companies with warrants and/or executive stock options. Mimeo.
Day, R. and Lewis, C. (1988). The behaviour of the volatility implicit in the prices of stock index options. *Journal of Financial Economics*, 22:103–122.
Day, R. and Lewis, C. (1992). Stock market volatility and the information content of stock index options. *Journal of Econometrics*, 52:267–287.
Derman, E. and Kani, I. (1994). Riding on the smile. *Risk*, 7:32–39.
Duan, J. (1995). The GARCH option pricing model. *Mathematical Finance*, 5(1):13–32.
Dumas, B., Fleming, J. and Whaley, R. (1998). Implied volatility functions. Empirical tests. *Journal of Finance*, 53:2059–2106.
Dupire, B. (1994). Pricing with a smile. *Risk*, 7, 18–20.
Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors – an empirical Bayes approach. *Journal of the American Statistical Association*, 68:117–130.
Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variances of UK inflation. *Econometrica*, 50:987–1008.
Engle, R. F. and Bollerslev, T. (1986). Modelling persistence of conditional variances. *Econometric Reviews*, 5:1–50.
Engle, R. F. and Mustafa, C. (1992). Implied ARCH models from option prices. *Journal of Econometrics*, 52:289–311.
Epps, T. and Epps, M. (1976). The stochastic dependence of security price changes and transactions volume. Implications for the mixture of distributions hypothesis. *Econometrica*, 44:305–321.
Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, Vol. II. New York, NY: John Wiley & Sons.
French, K. R., Schwert, G. W. and Stambaugh, R. (1987). Expected stock returns and volatility. *Journal of Financial Economics*, 19:3–29.
Garcia, R., Luger, R. and Renault, E. (2001). *Empirical Assessment of an Intertemporal Option Pricing Model with Latent Variables*. Montreal: CIRANO Scientific Series.
Garman, M. and Klass, M. (1980). On estimation of security price volatility from historical data. *Journal of Business*, 53:67–78.
Ghysels, E., Harvey, A. and Renault, E. (1996). Stochastic volatility. In: G. S. Maddala and C. R. Rao (eds), *Handbook of Statistics*, Vol. 14. Amsterdam: Elsevier.
Hull, J. and White, A. (1987). The pricing of options on assets with stochastic volatilities. *Journal of Finance*, 42:281–300.
Hutchinson, J., Lo, A. and Poggio, T. (1994). A nonparametric approach to the pricing and hedging of derivative securities via learning networks. *Journal of Finance*, 49:851–889.
Hwang, S. and Satchell, S. (1998). Implied volatility forecasting: a comparison of different procedures including fractionally integrated models with applications to UK equity options. In: J. Knight and S. Satchell (eds), *Forecasting Volatility in the Financial Markets*. Oxford: Butterworth–Heinemann.
Jackwerth, J. and Rubinstein, M. (1996). Recovering probability distributions from option prices. *Journal of Finance*, 51: 1611–1631.
Karolyi, G. A. (1993). A Bayesian approach to modelling stock return volatility for option valuation. *Journal of Financial and Quantitative Analysis*, 28:579–594.
Knight, J. and Satchell, S. (1997). Existence of unbiased estimators of the Black/Scholes option price, other derivatives, and hedge ratios. *Econometric Theory*, 13:791–807.
Latane, H. and Rendleman, R. (1976). Standard deviation of stock prices emplied in option prices. *Journal of Finance*, 31:369–381.
Longstaff, F. (1995). Option pricing and the Martingale restriction. *Review of Financial Studies*, 8(4):1091–1124.
Melino, A. and Turnbull, S. M. (1990). Pricing foreign currency options with stochastic volatility. *Journal of Econometrics*, 45:239–265.
Merton, R. (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science*, 4:141–183.
Morgan, I. (1976). Stock prices and heteroscedasticity. *Journal of Business*, 49:496–508.

Ncube, M. and Satchell, S. (1997). The statistical properties of the Black–Scholes option price. *Mathematical Finance*, 7:287–305.
Nelson, D. B. (1990). ARCH models as diffusion approximations. *Journal of Econometrics*, 45:7–38.
Nelson, D. B. (1991). Conditional heteroscedasticity in asset returns. A new approach. *Econometrica*, 59:347–370.
Noh, J., Engle, F. and Kane, A. (1994). Forecasting volatility and option prices of the S&P 500 index. *Journal of Derivatives*, 2: 17–30.
Parkinson, M. (1980). The extreme value method of estimating the variance of the rate of return. *Journal of Business*, 53:61–66.
Poterba, J. and Summers, L. (1987). The persistence of volatility and stock market fluctuations. *American Economic Review*, 76:1142–1151.
Rubinstein, M. (1994). Implied binomial trees. *Journal of Finance*, 49:771–818.
Scott, L. (1991). Random variance option pricing. *Advances in Futures and Options Research*, 5:113–135.
Tauchen, G. and Pitts, M. (1983). The price variability–volume relationship on speculative markets. *Econometrica*, 51:485–505.
Watson, G. (1944). *A Treatise on the Theory of Bessel Functions*, 2nd edn. Cambridge: Cambridge University Press.
Wiggins, J. (1987). Option values under stochastic volatility. Theory and empirical estimates. *Journal of Financial Economics*, 19:351–372.

7 Bayesian forecasting of options prices: a natural framework for pooling historical and implied volatility information[1]

Theo Darsinos and Stephen Satchell

Abstract

Bayesian statistical methods are naturally oriented towards pooling, in a rigorous way, information coming from separate sources. It has been suggested that both historical and implied volatilities convey information about future volatility. However, typically, in the literature, implied and return volatility series are fed separately into models to provide rival forecasts of volatility or options prices. We develop a formal Bayesian framework where we can merge the backward-looking information as represented in historical daily return data with the forward-looking information as represented in implied volatilities of reported options prices. We apply our theory in forecasting (in- and out- of sample) the prices of FTSE 100 European index options. We find that for forecasting option prices out of sample (i.e. 1 day ahead), our Bayesian estimators outperform standard forecasts that use implied or historical volatilities – particularly so when we assess the merit of the forecasts in economic or directional accuracy terms (i.e. in terms of the profit to a trading strategy).

7.1 Introduction

The purpose of this chapter is to propose a Bayesian approach for forecasting (in- and out- of sample) the prices of European options. In the classical Black–Scholes (1973) framework and its subsequent extensions, this basically requires estimating *ex ante* the volatility of financial asset returns. As Engle and Mustafa (1992) suggest, there are two approaches available to the analyst undertaking this task:

1. The direct one, although backward-looking in nature, is to use high frequency historical data on the behaviour of asset prices either to calculate some statistic, such as the standard deviation of returns, or to explicitly estimate the stochastic process of volatility via maximum likelihood or other methods. The former usually applies when a constant volatility option pricing model is used, while the latter is more relevant for a (discrete-time) stochastic volatility option pricing model.
2. The indirect one, first introduced by Latane and Rendleman (1976), is forward-looking in nature, and uses the market prices of traded options together with an option pricing model (e.g. the Black–Scholes) to infer expectations about future volatility. This is

[1]An earlier version of this chapter has been published under the same title as Working Paper No. 0116 in the Department of Applied Economics (DAE) Discussion Paper Series, University of Cambridge.

done by exploiting the monotonicity of the option price with respect to volatility to invert the option pricing formula in terms of the volatility parameter. It is the so-called implied volatility approach, which has proved very popular amongst market practitioners but has also helped uncover the limitations of the Black–Scholes model.

Indeed, it is now widely recognized that the Black–Scholes constant volatility assumption is no longer sufficient to capture modern market (i.e. post-1987 crash) phenomena (see, for example, Rubinstein (1994) for a discussion of the observed pattern of implied volatilities known as the 'smile' effect). Although there has been a lot of work done in modifying the specification of volatility to make it a stochastic process, there has not yet been a model of stochastic volatility that enjoys the popularity of Black–Scholes. This is partly due to the many challenges that arise under stochastic volatility. First of all, volatility is a 'hidden' process, and therefore the process parameters are hard to estimate in a continuous framework with only discrete observations. Moreover, stochastic volatility introduces a new source of randomness that cannot be hedged and typically induces market incompleteness.[2] This in turn implies that there is not a unique, arbitrage free price for a contingent claim, and further assumptions about investors' preferences (utility) need to be made to restore market completeness. Alternatively, to restore market completeness in a stochastic volatility framework one must undertake the (prone to large estimation errors) task of determining empirically the unobservable market price of volatility risk (volatility risk-premium).[3]

Due to the difficulties associated with the empirical application of stochastic volatility models (on top of the aforementioned, they often lead to intractable results or require cumbersome numerical analysis and lengthy simulations), financial practitioners, in as much as they announce what they do, seem to continue to use the Black–Scholes model, albeit in an *ad hoc* fashion. For example, they often use GARCH to predict volatility, and then use the traditional Black–Scholes coupled with GARCH to price the option. This hybrid procedure, whilst lacking theoretical rigour, can be partially justified by the arguments of Amin and Jarrow (1991), and by the empirical results of Baillie and Bollerslev (1992), Engle and Mustafa (1992), Satchell and Timmermann (1993) and Duan (1995). In this sort of framework, Noh *et al.* (1994) assess the effectiveness of ARCH models for pricing options. Their study compares predictions of S&P 500 index options prices from GARCH with predictions of the same options prices from forecasting implied volatility. They suggest that both methods can effectively forecast prices well enough to

[2] An exception occurs if one (unrealistically) assumes that the volatility process is either uncorrelated with the underlying (this in effect implies a market price of volatility risk equal to zero) or consider a model where the volatility and the underlying processes are perfectly correlated (see, for example, Hull and White, 1987 for a discussion).

[3] There are many widely cited papers on stochastic volatility option pricing – to name just a few, Hull and White (1987), Scott (1987), Wiggins (1987), Stein and Stein (1991), Heston (1993) and Hobson and Rogers (2000) address the issue in a continuous time framework, while Satchell and Timmermann (1993), Duan (1995), Heston and Nandi (2000) and Duan and Zhang (2001) attack the problem in discrete time. Very briefly, the former class of papers model the volatility process as a diffusion while the latter model volatility as a GARCH process. One of the advantages of modelling volatility as a GARCH stochastic process (as opposed to a purely stochastic process) is that no additional source of randomness is introduced. Interestingly, since the work of Nelson (for example, Nelson, 1990) there has been a lot of theoretical interest in the convergence of discrete-time heteroscedastic volatility models to continuous time stochastic volatility models, particularly since the former present the comparative advantage of ease of estimation of the process parameters.

profit by trading if transaction costs are not too high. However, they also claim that volatilities incorporated in option prices do not fully utilize historical information, and that GARCH volatility forecasts could add value.

Another *ad hoc* procedure often used by practitioners to deal with the (implied) volatility smile and term structure is to regress the past Black–Scholes implied volatility of an option to its strike prices and maturities (see Dumas *et al.*, 1998). This estimated relationship then serves as the basis for calibrating the future implied volatilities for options with different strikes and maturities. Indeed, discussions with market practitioners indicate that they are happy using the Black–Scholes model as long as it is adjusted for the volatility smile. In fact, using the Black–Scholes model with a strike-dependent volatility function is another way of expressing the belief that the underlying deviates from its log-normal behaviour. Harvey and Whaley (1992) use time-series regressions of option implied volatilities to forecast the 1 day ahead volatility of S&P 100 index options. They reject the null hypothesis that volatility changes are unpredictable on a daily basis. However, after accounting for transaction costs, a trading strategy based upon out-of-sample volatility forecasts does not generate abnormal returns. Day and Lewis (1992) introduce implied volatilities into a GARCH and EGARCH model, and find that they have some explanatory power for predicting variance in most models, but that in no case are they adequate for predicting implied volatilities.

One subclass of stochastic volatility models that enjoys popularity amongst practitioners is the class of what Rebonato (1999) calls 'restricted stochastic volatility models', otherwise commonly known as 'deterministic (level-dependent) volatility models'. These models describe the stochastic evolution of the state variable by means of a volatility term that is a deterministic function of the stochastic underlying stock price[4] (see, for example, Cox and Ross, 1976). The advantage of these models is that they preserve market completeness since the (stochastic) volatility functionally depends on the underlying. Assuming a 'restricted stochastic volatility model', Dupire (1994), Rubinstein (1994) and Derman and Kani (1998) provide tree-based algorithms to extract from observed option prices of different strikes and maturities a volatility function that is capable of fitting the cross-section of option prices (i.e. reproducing the smile). However, Dumas *et al.* (1998) test the predictive and hedging performance of these models and find that it is no better than the *ad hoc* procedure (discussed in the previous paragraph) that merely smooths Black–Scholes implied volatilities across exercise prices and times to expiration. This is interpreted as evidence that more complex (than the constant) volatility specifications overfit the observed structure of option prices.

The purpose of this chapter is to present a new Bayesian methodology that can potentially improve upon the aforementioned existing methods for forecasting options prices. In particular, our contribution is twofold:

1. Bayesian statistical methods are naturally oriented towards pooling, in a rigorous way, information coming from separate sources. It has been suggested that both historical and implied volatilities convey information about future volatility. However, typically,

[4]Or in other words, the volatility of the underlying σ should only exhibit the stochastic behaviour allowed by the functional dependence on the stock price S. Therefore, under a restricted volatility model the underlying process is given by: $dS(t) = \mu(S, t)dt + \sigma(S, t)dW_t$. This equation describes the most general set-up that goes beyond the case of a purely deterministic (time-dependent) volatility, and still allows risk-neutral valuation without introducing other hedging instruments apart from the underlying itself.

in the literature, implied and return volatilities series are fed separately into models to provide rival forecasts of volatility or options prices.[5] We develop a formal Bayesian framework where we can merge the backward-looking information, as represented in historical daily return data, with the forward-looking information, as represented in implied volatilities of reported options prices. In a recent paper (Darsinos and Satchell, 2001) we have derived the prior and posterior densities of the Black–Scholes option price. In this chapter, we extend our previous analysis from a modelling context to a forecasting context by deriving the predictive density of the Black–Scholes option price. We also apply our theory in forecasting the prices of FTSE 100 European index options. We find that our Bayesian forecasts are generally as good (in terms of Mean Mispricing Error, MME) as other benchmark forecasts that use implied/historical volatility for explaining the observed market prices of options (i.e. in-sample forecasting); are slightly better (in terms of Mean Forecasting Error, MFE) for forecasting option prices 1 day ahead (i.e. out-of-sample forecasting); and are substantially better if assessed in economic or directional accuracy terms (i.e. in terms of the out-of-sample profit to a trading strategy).

2. Forecasting options prices has typically been synonymous with forecasting volatility. In fact, attempting directly to forecast options prices instead of volatility might at first appear unusual. However, option prices are substantially influenced by the volatility of underlying asset prices as well as the price itself. The majority of existing theories of (out-of-sample) option price forecasting (e.g. Harvey and Whaley, 1992; Noh *et al.*, 1994) use only the implied volatility or historical volatility (e.g. GARCH) to forecast option prices while keeping the price of the underlying fixed. That is, the closing price of the underlying today is used as a forecast of tomorrow's value. In our Bayesian framework we treat both the underlying and its volatility as random variables, and the predictive density introduced in this chapter explicitly incorporates uncertainty in both price and volatility. Regarding the stochastic or non-stationary character of volatility, we are able by the very nature of our approach to introduce this parameter in a probabilistic rather than deterministic way. Bayesian statistics treat the parameters of distributions of random variables as random variables themselves, and assign to them probability distributions. This adds an important element of flexibility to our method.[6] For example, a Bayesian Black–Scholes framework is not as restrictive as a classical Black–Scholes framework, and can accommodate for non-constant volatility across strikes and maturities by incorporating an informative prior. In this chapter, for example, we use historical market data which, for a given day, provide a volatility estimate that is common for options of all strikes and maturities. However, we also pool this common market estimate with implied volatility data which, for a given day, are strike- and maturity-specific. The resulting posterior (and predictive) estimate of volatility therefore varies across strikes and maturities.

Bayesian methods have also been used in the past for the valuation of options. For example, Karolyi (1993) utilizes prior information extracted from the cross-sectional patterns

[5] For an exception, see Day and Lewis (1992). There the authors add the implied volatility as an exogenous variable to GARCH-type models to examine the incremental information content of implied volatilities.

[6] Indeed, as noted by Bauwens and Lubrano (2000), the predictive method, which incorporates an additional source of uncertainty, is a better alternative to using a marginal measure to be plugged in the Black–Scholes formula, which can be very dangerous – particularly at times of near non-stationarity.

in the return volatilities for groups of stocks sorted either by size or financial leverage or by trading volume, together with the sample information, to derive the posterior density of the variance. He reports improved prediction accuracy for estimates of option prices calculated using the Bayesian volatility estimates relative to those computed using implied volatility, standard historical volatility, or even the actual ex-post volatility that occurred during each option's life. More recently, Bauwens and Lubrano (2000) show how option prices can be evaluated from a Bayesian viewpoint using a GARCH model for the dynamics of the volatility of the underlying asset. Their methodology delivers (via a numerical algorithm) the predictive distribution of the payoff function of the underlying. The authors suggest that this predictive distribution can be utilized by market participants to compare the Bayesian predictions to realized market prices or to other predictions. Our chapter differs from Bauwens and Lubrano's in that we follow the log-normal Black–Scholes structure (which allows us to derive the posterior distribution of the option price in analytic form) whilst they follow a GARCH discrete-time structure similar to the one suggested in Duan (1995) or, more recently, in Hafner and Herwartz (1999).

The organization of the chapter is as follows: In Section 7.2 we outline the classical distributional assumptions behind the Black–Scholes model and its estimation. In Section 7.3 we work towards establishing a Bayesian option pricing framework and extend our previous work (Darsinos and Satchell, 2001) from a modelling context to a forecasting context. The posterior and predictive densities of the Black–Scholes option price are derived. Section 7.4 deals with the empirical implementation of our model. We test the predictive performance of our Bayesian distributions when applied to the market of FTSE 100 European index options. Concluding remarks follow in Section 7.5.

7.2 A classical framework for option pricing

We start with the classical Black–Scholes assumption that the stock price P_t follows a Geometric Brownian Motion. This then implies the following formula for the stock price:

$$P_t = P_0 \exp\left(\left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma W_t\right) \tag{7.1}$$

where W_t is a standard Brownian motion ($W_0 = 0$), P_0 is the initial price at time 0, μ is the instantaneous mean and σ^2 is the instantaneous variance. The Black–Scholes option price for a European call option is then given by

$$C_t = C_{BS}(P_t, \sigma) = P_t\Phi(d_1) - K\exp(-r\tau)\Phi(d_2) \tag{7.2}$$

where $d_1 = \dfrac{\log(P_t/K) + (r + \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}}$, $d_2 = d_1 - \sigma\sqrt{\tau}$ and $\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}\exp(-\frac{y^2}{2})dy$, K is the exercise price at the expiry date T, r is the risk-free rate of interest, and $\tau = T - t$ is the time to maturity. Note that the only unobservable parameter entering the valuation formula is the variance parameter σ^2. The next step in the valuation problem is therefore to estimate σ^2.

7.2.1 Historical information (sample information)

The classical minimum-variance unbiased estimator of σ^2 for t observations of (daily) continuously compounded returns is given by the sample variance $s^2 = \sum_{i=1}^{t} (x_i - \bar{x})^2/(t-1)$ where x is the log-return between two consecutive time intervals (i.e. $x_j = \log(P_j/P_{j-1})$) and $\bar{x}$ is the sample mean return (i.e. $\bar{x} = (1/t)\sum_{j=1}^{t} x_j$). It is well known that the statistic $(t-1)s^2/\sigma^2$ has a χ^2 (chi-square) distribution with $t-1$ degrees of freedom.[7] From this we can obtain the probability density function of the sample variance (or the likelihood function of the true variance). It is given by:

$$f(s^2 \backslash \sigma^2, t) \equiv L(\sigma^2 \backslash s^2, t) \equiv \left(\frac{t-1}{2}\right)^{\frac{t-1}{2}} \frac{(s^2)^{\frac{(t-1)}{2}-1}}{\Gamma\left(\frac{t-1}{2}\right)(\sigma^2)^{\frac{t-1}{2}}} \exp\left(-\frac{(t-1)s^2}{2\sigma^2}\right) \tag{7.3}$$

where $L(\,)$ denotes a likelihood function.

Since we want to work with standard deviations rather than variances, we transform equation (7.3) to represent the distribution of the sample standard deviation:

$$f(s \backslash \sigma, t) = 2\left(\frac{t-1}{2}\right)^{\frac{t-1}{2}} \frac{s^{t-2}}{\Gamma\left(\frac{t-1}{2}\right)\sigma^{t-1}} \exp\left(-\frac{(t-1)s^2}{2\sigma^2}\right) \tag{7.4}$$

Note that here and throughout we will use $f(\,)$ to denote probability density functions generally, and not one specific probability density. The argument of $f(\,)$ as well as the context in which it is used will identify the particular probability density being considered.

Also from equation (7.1) we have that the stock price P_t is log-normally distributed. Its probability density function is given by:

$$f(P_t \backslash \mu, \sigma, t) = \frac{1}{P_t\sqrt{2\pi t}\sigma} \exp\left\{\frac{-\left[\ln\left(\frac{P_t}{P_0}\right) - \left(\mu - \frac{\sigma^2}{2}\right)t\right]^2}{2\sigma^2 t}\right\} \tag{7.5}$$

Note that for notational simplicity we will ignore dependence on P_0.

7.3 A Bayesian framework for option pricing

In the classical Black–Scholes framework, the drift and diffusion parameters are regarded as constants. In a Bayesian framework, these parameters are introduced in a probabilistic rather than deterministic way and are treated as random variables. We should therefore

[7] The probability density function of a variable z that is distributed chi-squared with $t-1$ degrees of freedom is given by $f(z) = \frac{1}{\Gamma((t-1)/2)2^{(t-1)/2}} z^{((t-1)/2)-1} \exp(-z/2)$.

identify probability distributions for the drift and diffusion parameters. In identifying these distributions we follow standard Bayesian methodology, as presented in Raiffa and Schlaifer (1961), Zellner (1971) and, more recently, Hamilton (1994) and Bauwens *et al.* (1999).

Regarding now the choice of measure, we note that options are priced under the risk-neutral measure and the resulting option prices are independent of the drift of the underlying. For forecasting purposes, however, the objective measure should be used, since the predictive density is not invariant on the drift. In other words, although the risk-neutral measure is appropriate for asset pricing, for calculating objective probabilities such as (for example) the option finishing in the money, the objective measure is appropriate. In the rest of this chapter the derived distributions are with respect to the objective measure. For the risk-neutral posterior distribution of the Black–Scholes option price, see Darsinos and Satchell (2001).

7.3.1 Drift information (non-informative prior)

Following Darsinos and Satchell (2001), we have that the conditional probability density function of the expected rate of return μ is given by:

$$f(\mu \backslash \sigma, t, m) = \frac{\sqrt{t}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{t(\mu - m)^2}{2\sigma^2}\right) \tag{7.6}$$

where m is a hyperparameter. In calculating m we intend to use the 'empirical' Bayes approach. That is, we estimate the hyperparameter from the sample information $(1, 2, \ldots, t)$. This can be viewed as incorporating a non-informative prior for the expected rate of return of the underlying. Alternatively, one could use prior sample information or information in the form of analysts forecasts.

7.3.2 Implied volatility information (informative prior)

As a source of prior information we use the implied volatilities or the at-the-money implied volatilities of reported option prices. We assume that the variance has an inverted-gamma-1 distribution with prior parameters $(t's'^2/2, t'/2)$.

$$f(\sigma^2 \backslash s'^2, t') = \left(\frac{t's'^2}{2}\right)^{\frac{t'}{2}} \frac{\exp\left(-\frac{t's'^2}{2\sigma^2}\right)}{\Gamma\left(\frac{t'}{2}\right)(\sigma^2)^{\frac{t'}{2}+1}} \tag{7.7}$$[8]

[8] *Remark*: when the distributions are conditional on any prior parameters (i.e. s', t' and m) and on t (it is not unreasonable to assume that the sample size is known before the sample is drawn), we will refer to these distributions as prior or unconditional.

Transforming the above equation to represent the distribution of volatility, we have:

$$f(\sigma \backslash s', t') = 2\left(\frac{t's'^2}{2}\right)^{\frac{t'}{2}} \frac{\exp\left(-\frac{t's'^2}{2\sigma^2}\right)}{\Gamma\left(\frac{t'}{2}\right)(\sigma)^{t'+1}} \tag{7.8}$$

Note that s' represents the implied volatility estimate, and t' the weight attached to it. For example, the analyst might use a month of implied volatility data and calculate s' as the sample mean of this data (alternatively, s' can represent a composite measure of implied volatilities). The weight chosen to be attached to s' can then be, for example, a month ($t' = 30$). However, this may not necessarily be the case, and the weight can be whatever the analyst feels is appropriate. To illustrate, consider the example where the previous day's implied volatility estimate is used as prior information. Here, if $t' = 1$ (i.e. 1 day) is chosen, the prior (implied volatility) information will be absorbed by the sample information since if we assume that the analyst used 1 month (say $t = 30$ days) of daily historical returns to estimate the sample standard deviation s, the weight attached to the sample information will be $t - 1 = 29$ (1 degree of freedom is lost in estimating s), while the weight attached to the implied volatility information will be just 1. To sum up, the analyst can, if believing strongly in the prior information, use as much weight as he or she feels it merits. This can be a powerful tool in the hand of the analyst, since different point (or interval) estimates can be obtained using different weights. Bayesian methods are in a way both a science and an art!

7.3.3 The posterior density of the Black–Scholes option price

In contrast to classical analysis, where the main piece of output is a point estimate, Bayesian analysis produces as its main piece of output the so-called posterior density. This posterior density can then be combined with a loss or utility function to allow a decision to be made on the basis of minimizing expected loss or maximizing expected utility. For example, for positive definite quadratic loss functions, the mean of the posterior distribution is an optimal point estimate. If the loss is proportional to the absolute value of the difference between the true and the estimated values then the median is chosen, while a zero loss for a correct estimate and a constant loss for an incorrect estimate leads to the choice of the mode.

We now illustrate, in three steps, how we can derive the posterior density of the Black–Scholes option price by using equations (7.2)–(7.8) above. Note that we will use only symbolic notation. The analytical formulae for all the densities involved in the calculations are exhibited for reference in the Appendix to this chapter. For their derivations, see Darsinos and Satchell (2001).

Since the option price as an unconditional random variable depends both on the underlying and its volatility, we must (1) obtain the posterior density of price and volatility, then the posterior density of the option price follows after (2) applying a non-linear transformation, and (3) dividing by the conditional (on the sample and prior information) density of the asset price.

1. We start from the densities of the drift (i.e. equation (7.6)) and of volatility (i.e. equation (6.8)). Then the joint density of drift and volatility is given by:

$$f(\mu, \sigma \backslash t, s', t', m) = f(\mu \backslash \sigma, s', t', m) f(\sigma \backslash s', t', (m)) \tag{7.9}$$[9]

Now, using equation (7.9) and the distribution of the underlying (equation (7.5)), we get:

$$f(P_t, \mu, \sigma \backslash t, s', t', m) = f(\mu, \sigma \backslash t, s', t', m) f(P_t \backslash \mu, \sigma, t, (s'), (t'), (m)) \tag{7.10}$$

Integrating out the drift rate, we get the 'prior' density of price and volatility (see footnote 8):

$$f(P_t, \sigma \backslash t, s', t', m) = \int_{-\infty}^{\infty} f(P_t, \mu, \sigma \backslash t, s', t', m) d\mu \tag{7.11}$$

Then, applying Bayes rule, the posterior density of price and volatility is given by:

$$\begin{aligned} f(P_t, \sigma \backslash s, t, s', t', m) &= \frac{f(P_t, \sigma \backslash t, s', t', m) f(s \backslash (P_t), \sigma, t, (s'), (t'), (m))}{f(s \backslash t, s', t', (m))} \\ &= \frac{f(P_t, \sigma, s \backslash t, s', t', m)}{f(s \backslash t, s', t', (m))} \end{aligned} \tag{7.12}$$

Observe that the numerator of the above equation – i.e. $f(P_t, \sigma, s \backslash t, s', t', m)$ – is readily obtained by multiplying equations (7.11) and (7.4), while the denominator – i.e. $f(s \backslash t, s', t', (m))$ – is derived from the following calculations:

Multiplying equation (7.7) with equation (7.3) we get

$$f(s^2, \sigma^2 \backslash t, s'^2, t') = f(\sigma^2 \backslash (t), s'^2, t') f(s^2 \backslash \sigma^2, t, (s'^2), (t')) \tag{7.13}$$

Then $f(s^2 \backslash t, s', t') = \int_0^{\infty} f(s^2, \sigma^2 \backslash t, s'^2, t') d\sigma^2$. Finally

$$f(s \backslash t, s', t', (m)) = 2sf(s^2 \backslash t, s'^2, t') \tag{7.14}$$

2. Having obtained the posterior density $f(P_t, \sigma \backslash s, t, s', t', m)$ and remembering that $C_t = C_{BS}(P_t, \sigma)$ (defined in equation (7.2)), we now apply the non-linear transformation:

$$P_t = P_t$$

$$\sigma = C_{BS}^{-1}(C_t) \equiv C_{BS}^{-1}(P_t, C_t) \Leftrightarrow C_t = C_{BS}(P_t, \sigma)$$

We invert the Black–Scholes option pricing formula in terms of σ for fixed P_t, thus obtaining σ as a function of C_t and P_t. This is the so-called implied volatility of the

[9] Observe that in the density of volatility $f(\sigma \backslash s', t', (m))$, (m) appears in parenthesis. Here and below, when a parameter is exhibited in parenthesis we take this to mean that it does not actually appear in the analytic formula for that specific density but, for coherence of the argument, we include it in the symbolic notation.

option price, and there is known to be a unique (one-to-one) inverse function from the monotonicity of the option price as a function of volatility. Note, however, that there is no analytic expression (with the exception of an at-the-money option) for $\sigma = C_{BS}^{-1}(P_t, C_t)$, and it will have to be evaluated numerically using a Newton–Raphson iterative procedure.

Applying the transformation, we get

$$f(P_t, C_t \backslash s, t, s', t', m) = f(P_t, \sigma = C_{BS}^{-1}(P_t, C_t) \backslash s, t, s', t', m)\,|J| \tag{7.15}$$

where J is the Jacobian of the non-linear transformation and is given by:

$$\frac{1}{J} = \begin{vmatrix} \partial P_t/\partial P_t & \partial P_t/d\sigma \\ \partial C_t/\partial P_t & \partial C_t/\partial \sigma \end{vmatrix} = \frac{\partial C_t}{\partial \sigma} = Vega$$

where $Vega = \phi\left(\dfrac{\ln\left(\dfrac{P_t}{Ke^{-r\tau}}\right) + \dfrac{C_{BS}^{-1}(P_t, C_t)^2\tau}{2}}{C_{BS}^{-1}(P_t, C_t)\sqrt{\tau}}\right) P_t\sqrt{\tau}$

$\phi(\ldots) = \Phi'(\ldots)$ denotes the standard normal probability density function.

3. Finally, to obtain the posterior density of the Black–Scholes option price we divide the joint density of price and option price (i.e. equation (7.15)) with the conditional (on the sample information) density of the underlying price:

$$f(C_t \backslash P_t, s, t, s', t', m) = \frac{f(P_t, C_t \backslash s, t, s', t', m)}{f(P_t \backslash s, t, s', t', m)} \tag{7.16}$$

From Darsinos and Satchell (2001), the analytic expression for the posterior density of the Black–Scholes option price is given by:

$$f(C_t \backslash P_t, s, t, s', t', m)$$

$$= \frac{|J|\left(\dfrac{2(t's'^2 + (t-1)s^2)t + (\ln(P_t/P_0) - mt)^2}{64}\right)^{\frac{t'+t}{4}}}{K_{\frac{t'+t}{2}}\left(\dfrac{1}{4}\sqrt{2(t's'^2 + (t-1)s^2)t + \left(\ln\left(\dfrac{P_t}{P_0}\right) - mt\right)^2}\right) C_{BS}^{-1}(P_t, C_t)^{t'+t+1}\left(\dfrac{t}{16}\right)^{\frac{t'+t}{2}}}$$

$$\times \exp\left(-\frac{t's'^2 + (t-1)s^2}{2C_{BS}^{-1}(P_t, C_t)^2} + \frac{\ln(P_t/P_0) - mt}{4}\right)$$

$$\times \exp\left(-\frac{1}{4C_{BS}^{-1}(P_t, C_t)^2 t}\left[\ln\left(\frac{P_t}{P_0}\right) - \left(m - \frac{1}{2}C_{BS}^{-1}(P_t, C_t)^2\right)t\right]^2\right)$$

where $K_{\frac{t'+t}{2}}(\ldots)$ represents the modified Bessel function of the second kind of order $(t'+t)/2$.

Although seemingly complicated, the above density is in fact very simple and fully operational. We subject it to an immediate numerical test to verify that it is a proper density. We are able to confirm that it does integrate to one.[10]

A number of different point or interval estimators can be obtained from the above density. As a point estimator, the most popular is probably the mean of the posterior density, which is optimal under squared error loss. However, as already mentioned above, the median and the mode can also prove useful point estimators. The derived density is also ideal for quantile estimation and value-at-risk (VaR) calculations (the return distribution for a call option is obtained by taking the $\log C_t/C_{t-1}$ transform of the distribution of equation (7.16)). Darsinos and Satchell (2001) show that the posterior probability distribution for a long call option generally exhibits excess kurtosis and is positively skewed. In particular, for an at-the money option the distribution is close to normal; however, as we move progressively out- or in-the-money the distribution of the option price exhibits an increasingly thinner left tail than the normal distribution. The VaR for an asset or portfolio of assets is critically dependent on the left tail of their distributions. If, for example, one assumes that the return distribution of a long call option is normal, one will tend to calculate a VaR that is higher than the true VaR. Similarly, for a short call the calculated VaR will be too low.

7.3.4 The predictive density of the Black–Scholes option price

We now extend the work of Darsinos and Satchell (2001) from a modelling context to a forecasting context. On many occasions, given our sample information, we are interested in making inferences about other observations that are still unobserved – one part of the problem of prediction. In the Bayesian approach, the probability density function for the as yet unobserved observations given our sample information can be obtained, and is known in the Bayesian literature as the predictive density.

In our case, given our sample information (P_t, s) (note that s is computed from $P_0 \ldots P_t$), we are interested in making inferences about future option values $C_T = C(P_T, \sigma)$ for some $T > t$. To obtain the predictive density of the option price, we proceed as follows. In equation (7.12) above we have shown how to derive the posterior density of price and volatility given the sample and prior information. In analytic form, it is given by:

$$f(P_t, \sigma \backslash P_0, s, t, s', t', m) = \frac{\left(\frac{t's'^2+(t-1)s^2}{2}\right)^{\frac{t'+t-1}{2}}}{\sqrt{\pi t} P_t \Gamma\left(\frac{t'+t-1}{2}\right)} \frac{1}{\sigma^{t'+t+1}} \exp\left(-\frac{t's'^2+(t-1)s^2}{2\sigma^2}\right)$$
$$\times \exp\left(-\frac{1}{4\sigma^2 t}\left[\ln\left(\frac{P_t}{P_0}\right) - \left(m - \frac{1}{2}\sigma^2\right)t\right]^2\right)$$

[10] Just to report a set of trial values that we used (in daily format):

$t' = 30, s' = 0.010174, t = 30, s = 0.0076, m = 0.0005,$

$P_t = 3157, P_0 = 3148, K = 3025, r = 0.00022, \tau = 18.$

Without loss of generality, we can rewrite this as the joint distribution of the yet unobserved underlying price and of volatility, given the sample and prior information:

$$f(P_T, \sigma \backslash P_t, s, T, t, s', t', m) = \frac{\left(\frac{t's'^2+(t-1)s^2}{2}\right)^{\frac{t'+t-1}{2}}}{\sqrt{\pi(T-t)}P_T\Gamma\left(\frac{t'+t-1}{2}\right)} \frac{1}{\sigma^{t'+t+1}} \exp\left(-\frac{t's'^2+(t-1)s^2}{2\sigma^2}\right) \times \exp\left(-\frac{1}{4\sigma^2(T-t)}\left[\ln\left(\frac{P_T}{P_t}\right)-\left(m-\frac{1}{2}\sigma^2\right)(T-t)\right]^2\right) \tag{7.17}$$

Our next step is to derive the joint distribution of the unobserved future option price and future stock price at time T. Again remembering that $C_T = C_{BS}(P_T, \sigma)$, take $f(P_T, \sigma \backslash P_t, s, T, t, s', t', m)$ and consider the transformation:

$$P_T = P_T$$
$$\sigma = C_{BS}^{-1}(C_T) \equiv C_{BS}^{-1}(P_T, C_T)$$

with Jacobian:

$$1/J^* = \begin{vmatrix} \partial P_T/\partial P_T & \partial P_T/d\sigma \\ \partial C_T/\partial P_T & \partial C_T/\partial \sigma \end{vmatrix} = \phi\left(\frac{\ln\left(\frac{P_T}{Ke^{-r(\tau-(T-t))}}\right)+\frac{[C_{BS}^{-1}(P_T,C_T)]^2(\tau-(T-t))}{2}}{C_{BS}^{-1}(P_T,C_T)\sqrt{\tau-(T-t)}}\right) P_T\sqrt{\tau-(T-t)}$$

Then

$$f(P_T, C_T \backslash P_t, s, T, t, s', t', m) = f\left(P_T, \sigma = C_{BS}^{-1}(P_T, C_T) \backslash P_t, s, T, t, s', t', m\right)|J^*| \tag{7.18}$$

Finally, to obtain the predictive density of the option price we require a single integration:

$$f(C_T \backslash P_t, s, T, t, s', t', m) = \int f(P_T, C_T \backslash P_t, s, T, t, s', t', m)\partial P_T$$
$$= \frac{\left(\frac{t's'^2+(t-1)s^2}{2}\right)^{\frac{t'+t-1}{2}}}{\sqrt{\pi(T-t)}\Gamma\left(\frac{t'+t-1}{2}\right)} \times \int_0^\infty \frac{|J^*|\exp\left(-\frac{t's'^2+(t-1)s^2}{2[C_{BS}^{-1}(P_T,C_T)]^2}-\frac{1}{4[C_{BS}^{-1}(P_T,C_T)]^2(T-t)}\left[\ln\left(\frac{P_T}{P_t}\right)-\left(m-\frac{1}{2}\left[C_{BS}^{-1}\left(P_T,C_T\right)\right]^2\right)\left(T-t\right)\right]^2\right)}{\left[C_{BS}^{-1}\left(P_T,C_T\right)\right]^{t'+t+1}P_T}\partial P_T \tag{7.19}$$

In this case as well we are able to confirm that this is a proper density. Regarding the numerical evaluation procedure that could be applied for the calculation of the predictive density, see again Darsinos and Satchell (2001). Here it suffices to note that $C_{BS}^{-1}(P_T, C_T)$ is a $n \times m$ matrix of implied volatilities. (We evaluate $C_{BS}^{-1}(P_i, C_j)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$ spanning the whole range of attainable values of P_T and C_T, thus generating a $n \times m$ matrix of implied volatilities.)

The estimation of both the posterior (i.e. equation (7.16)) and predictive (i.e. equation (7.19)) distributions is relatively simple. Both densities can be evaluated solely on the basis of observables. Estimation of the parameters of the distributions requires only historical data on the underlying and reported option price data. In the forthcoming section we illustrate how the derived densities can be used for explaining and forecasting the market prices of FTSE 100 index European call options.

7.4 Empirical implementation

We aim to forecast call option prices both in- and out- of sample. This effectively means that in the former case we will use the posterior density of Section 7.3.3 to *explain* the observed market prices of call options and compare the precision of the Bayesian estimates with some benchmark forecasts that use historical or implied volatility. In the latter case we aim to *forecast* the prices of call option prices 1 day in the future using the predictive density of Section 7.3.4 above.

We use daily data from the London International Financial Futures and Options Exchange (LIFFE) for the period from September 1992 to December 2000. Our data concern FTSE 100 index European call option contracts. The data record for each contract contains the price of the underlying, the exercise price, the expiration date, the settlement price, the trading volume, the corresponding implied volatility and the at-the-money implied volatility. To proxy for the risk-free rate, the rate on a UK Treasury-bill of comparable maturity is used. Note that the underlying on the European contract is the price of the corresponding index future. One slight modification that therefore has to be made is that instead of the Black–Scholes model of equation (7.2) we will use Black's (1976) model for call options on futures:

$$C_t = e^{-r\tau}[F_t \Phi(d_1) - K\Phi(d_2)] \tag{7.20}$$

where $d_1 = \dfrac{\log(F_t/K) + \sigma^2\tau/2}{\sigma\sqrt{\tau}}$ and $d_2 = d_1 - \sigma\sqrt{\tau}$. F_t represents the corresponding future's price. (To obtain the Bayesian distributions for European puts one should, of course, use the respective formula for put options.)

We are aware that on data where there is little or no volume, the exchange uses artificially generated prices based on a system called Autoquote, which effectively uses Black's formula. To minimize such an effect we limit our range of investigation to option maturities ranging from one week to seven weeks. For maturities within this range there is always a reasonable volume of trading. Similarly, since for each contract we have a variety of exercise prices, we keep the contracts where the option was, at some stage during the period of investigation, close to the money. Our remaining data represent a variety of options out-of-, at- and in-the-money with maturities varying from one week to seven weeks.

7.4.1 In-sample forecasting

This section assesses the extent to which the Bayesian call option value estimators improve upon standard classical procedures in describing actual market prices of call options. By standard procedures we mean forecasts based on Black's model for options on futures (i.e. equation (7.20)). We use five different estimates of volatility to be plugged in an *ad hoc* fashion into Black's model, and thus provide five benchmark estimates of the price of the option today. In particular, as an estimate of volatility we use: (1) a 15-day moving average of historical volatilities (i.e. the sample standard deviation from daily returns over the 15 days preceding the date of each option price reported); (2) a 30-day moving average of historical volatilities; (3) a 15-day moving average of implied volatilities preceding the date of each option price; (4) a 30-day moving average of implied volatilities preceding the date of each option price reported; and (5) the previous day's implied volatility value.

For the Bayesian estimators, we take the mean of the posterior distribution as a point estimate. This is optimal under quadratic loss. We use four different Bayesian estimates. These are distinct in the sense that we use two different estimation horizons when estimating the parameters of the distribution (i.e. t, s, t', s', m). In particular, two estimates are derived using a 15-day estimation horizon and the other two using a 30-day estimation horizon. This effectively means that in the former case t and t' are both equal to 15 (i.e. we use 15 days of prior and sample information) while in the latter case they are equal to 30 (i.e. we use 30 days of prior and sample information). The parameter s is estimated as the sample standard deviation from daily returns over the 15 and 30 days respectively preceding the date of each option price reported. The parameter m is estimated as the mean of daily returns over the last 15 and 30 days respectively preceding each reported option price.

The two estimates that belong to the same estimation horizon are in turn distinct, since we use either the implied volatility or the at-the-money implied volatility to estimate s'. Hence s' is estimated as the mean of implied volatilities or the mean of at-the-money implied volatilities over the specified time horizon preceding the reported option price.

We measure overall fit to market data in terms of Mean Mispricing Error (MME). The Mispricing Error (ME) of each estimate is computed by $(\hat{C}_{it} - C_{it})/C_{it}$, where C_{it} is the market call option price and $\hat{C}_{it}$ is the estimated Bayesian or other (benchmark) option value. Then the MME is given by:

$$MME = (1/N) \sum_{i=1}^{N} (\hat{C}_{it} - C_{it})/C_{it} \tag{7.21}$$

We also assess the relative mispricing error (RME) of the option price estimates with respect to the time to maturity of the option and the degree to which the option is in- or out- of the money (moneyness). This is done by averaging the ME within the different subgroups. Note that the 'moneyness' of an option is measured as $(F_t/K) - 1$, where F_t is the underlying FTSE 100 index future price and K the exercise price of the option. Hence a negative value indicates an out-of-the-money option and a positive value an in-the-money option.

We use four randomly/arbitrarily selected European FTSE 100 index call option contracts, namely the June 1998, March 1999, September 2000 and December 2000 contracts. The moneyness of options in these contracts ranged from 5% out of the money to 8% in the money. The maturities considered were one to seven weeks. Table 7.1 summarizes our results.

Even before this exercise was undertaken, the evidence that implied volatilities fit option prices much better than historical volatilities was overwhelming. Our results from Table 7.1 confirm the aforementioned statement. The MME for the historical volatility estimates ranged between 10.8% and 11.4%, against a range of 4.2% and 6.6% for the

Table 7.1a Mean mispricing error of Bayesian and other call option price estimates

	Mean Mispricing Error (MME)
Benchmark models:	
Historical volatility (15-day moving average)	0.114
Historical volatility (30-day moving average)	0.108
Implied volatility (15-day moving average)	0.060
Implied volatility (30-day moving average)	0.066
Implied volatility (previous day)	0.042
15-day estimation horizon	
Bayesian (implied volatility – historical volatility)	0.050
Bayesian (A-T-M implied volatility – historical volatility)	0.060
30-day estimation horizon	
Bayesian (implied volatility – historical volatility)	0.053
Bayesian (A-T-M implied volatility – historical volatility)	0.072

Table 7.1b Relative mispricing errors of Bayesian and other call option price estimates with different times to maturity

Time to maturity:	**1 week**	**2 weeks**	**3 weeks**	**4 weeks**	**7 weeks**
Benchmark models:					
Historical volatility (15-day moving average)	0.088	0.028	0.122	0.218	0.112
Historical volatility (30-day moving average)	0.135	0.093	0.097	0.073	0.140
Implied volatility (15-day moving average)	0.108	0.057	0.026	0.081	0.030
Implied volatility (30-day moving average)	0.063	0.076	0.031	0.115	0.043
Implied volatility (previous day)	0.098	0.028	0.026	0.027	0.032
15-day estimation horizon					
Bayesian (implied volatility – historical volatility)	0.092	0.008	0.054	0.040	0.055
Bayesian (A-T-M implied volatility – historical volatility)	0.111	0.023	0.061	0.040	0.065
30-day estimation horizon					
Bayesian (implied volatility – historical volatility)	0.074	0.049	0.026	0.051	0.067
Bayesian (A-T-M implied volatility – historical volatility)	0.089	0.046	0.025	0.054	0.072

Table 7.1c Relative mispricing errors of Bayesian and other call option price estimates with different degrees of moneyness

Moneyness (M)	−5% < M < −2%	−2% < M < 0%	0% < M < 3%	3% < M < 8%
Benchmark models:				
Historical volatility (15-day moving average)	0.116	0.124	0.169	0.048
Historical volatility (30-day moving average)	0.253	0.086	0.112	0.049
Implied volatility (15-day moving average)	0.188	0.063	0.041	0.016
Implied volatility (30-day moving average)	0.116	0.068	0.081	0.023
Implied volatility (previous day)	0.139	0.050	0.021	0.008
15-day estimation horizon				
Bayesian (implied volatility – historical volatility)	0.121	0.050	0.049	0.015
Bayesian (A-T-M implied volatility – historical volatility)	0.161	0.051	0.054	0.019
30-day estimation horizon				
Bayesian (implied volatility – historical volatility)	0.145	0.052	0.048	0.013
Bayesian (A-T-M implied volatility – historical volatility)	0.156	0.053	0.050	0.018

implied volatility estimates. The estimate that used the reported implied volatility value of the previous day to be plugged into Black's model outperformed all other estimates with a MME of 4.2%. The performance of our Bayesian estimates paralleled that of the implied volatility estimates with MME ranging between 5.0%–7.2%. However, we cannot claim that the Bayesian method is superior in explaining the observed market prices of FTSE 100 European call options. We might have achieved better results had we used the previous day's reported implied volatility as prior information rather than the sample mean over the 15 or 30 days preceding the date of each reported option price. After all, it turned out that the previous day's implied volatility conveyed enough information to outperform the other estimates.[11]

7.4.2 Out-of-sample forecasting

Implied volatilities by definition perform very well in explaining the observed market prices of options. For example, analysts or even exchanges often calculate implied volatilities from actively traded options on a certain stock and use them to calculate the price of a less actively traded option on the same stock. The evidence, however, is not clear whether

[11] Other potential approaches that might have yielded better results could include using solely implied volatilities as a source of prior and sample information.

implied volatilities can on their own provide adequate forecasts of future volatility or indeed option prices.

Hence we now turn to the more interesting exercise of forecasting the prices of options 1 day ahead. For this exercise we assume that there are five agents, each following a particular forecasting method to predict the price of the FTSE 100 index European call of tomorrow. In particular, Agents 1, 2 and 3 use Black's (1976) model for options on futures. As an estimate of volatility to be plugged into the model, Agent 1 uses the sample standard deviation from daily returns over the 30 days preceding the date of each option price being reported; Agent 2 uses the mean of the implied volatilities over the 30 days preceding the date of each option price reported; and Agent 3 uses today's reported implied volatility value. As a forecast for tomorrow's value for the underlying, all three agents use today's price. It should be noted that the approach of Agent 3 is quite popular amongst market practitioners. Gemmill and Saflekos (2000) estimate the implied distribution for stock index options in London as a mixture of two log-normals and find that this method is somewhat better that the Black–Scholes (1-log-normal) approach at predicting out-of-sample option prices. However, according to the authors, an *ad hoc* model in which today's implied volatilities are applied to tomorrow's options does even better.

Agents 4 and 5 use the Bayesian predictive density of Section 7.3.4. They use the sample standard deviation from daily returns over the 30 days preceding the date of each option price reported to estimate the sample parameter s of the distribution. However, to estimate the prior parameter s' of the distribution, Agent 4 uses the mean of the implied volatilities over the 30 days preceding the date of each option price reported while Agent 5 uses the mean of the *at-the-money* implied volatilities over the 30 days preceding the date of each option price reported. Finally, the parameter m is estimated by both agents as the mean of daily returns over the last 30 days preceding each reported option price. As a point estimate of the option price, the agents take the mean of the predictive distribution.

We use twelve randomly selected European FTSE 100 index call option contracts, namely the September 1992, June 1993, December 1994, December 1995, December 1996, June 1997, June 1998, December 1998, March 1999, June 1999, September 2000 and December 2000 contracts. Then for *each* contract we fix eight dates for which we want to forecast the price of the option. Specifically, we obtain forecasts for the value of the option 1 week, 2 weeks, 18 days, 3 weeks, 25 days, 4 weeks, 30 days and 7 weeks before maturity. This effectively means that (for each contract) agents each apply their forecasting rule 8 days, 15 days, 19 days, 22 days, 26 days, 29 days, 31 days and 50 days respectively before maturity. Note here that the above dates were pre-specified arbitrarily/randomly.

In this exercise we measure overall forecasting performance in terms of the mean forecasting error (MFE):

$$MFE = (1/N) \sum_{i=1}^{N} (\hat{C}_{i(t+1)} - C_{i(t+1)})/C_{i(t+1)} \tag{7.22}$$

We also report the relative forecasting error (RFE) of the option price forecasts with respect to the time to maturity of the options. Our results are exhibited in Table 7.2.

This time the performance gap between the estimates that use either implied or historical volatilities is significantly reduced. Agent 1, who uses the historical volatility estimate, still produces the poorest forecasts (with an MFE of 23.1%), closely followed by Agents 2 and 3, who use the two implied volatility forecasts (with an MFE of 22.5% for Agent 2 and

Table 7.2a One-day ahead mean forecast error of Bayesian and other call option price estimates

	Mean Forecast Error (MFE)
Benchmark models:	
Agent 1: (historical volatility – 30-day moving average)	0.231
Agent 2: (implied volatility – 30-day moving average)	0.225
Agent 3: (today's implied volatility)	0.220
30-day estimation horizon	
Agent 4: (Bayesian – implied volatility – historical volatility)	0.196
Agent 5: (Bayesian – A-T-M implied volatility – historical volatility)	0.198

Table 7.2b One-day ahead relative forecast error (RFE) of Bayesian and other call option price estimates with different times to maturity

Time to maturity:	**1 week**	**2 weeks**	**18 days**	**3 weeks**
Benchmark models:				
Agent 1: (historical volatility – 30-day moving average)	0.311	0.370	0.228	0.209
Agent 2: (implied volatility – 30-day moving average)	0.406	0.351	0.257	0.167
Agent 3: (today's implied volatility)	0.443	0.340	0.195	0.186
30-day estimation horizon				
Agent 4: (Bayesian – implied volatility – historical volatility)	0.197	0.323	0.256	0.173
Agent 5: (Bayesian – A-T-M implied volatility – historical volatility)	0.194	0.321	0.258	0.174
Time to maturity:	**25 days**	**4 weeks**	**30 days**	**7 weeks**
Benchmark models:				
Agent 1: (historical volatility – 30-day moving average)	0.197	0.094	0.208	0.234
Agent 2: (implied volatility – 30-day moving average.)	0.140	0.133	0.193	0.155
Agent 3: (today's implied volatility)	0.161	0.095	0.169	0.168
30-day estimation horizon				
Agent 4: (Bayesian – implied volatility – historical volatility)	0.154	0.124	0.167	0.173
Agent 5: (Bayesian – A-T-M implied volatility – historical volatility)	0.156	0.138	0.162	0.180

22.0% for Agent 3). Agents 4 and 5, who use the Bayesian predictive density forecasts, outperform all others with the lowest MFE, at 19.6% and 19.8% respectively. In terms of the relative forecast error (RFE), we observe that the Bayesian forecasts dramatically outperform all other forecasts for close-to-maturity options (i.e. 1 week) and generally are as good as, or better, than the benchmark models.

To further assess the performance of the five agents we now devise the following simple trading rule. Our trading rule is similar in spirit to Noh *et al.* (1994). As mentioned above, during the sample period (September 1992 to December 2000), at the pre-specified dates, agents each apply their own forecasting rule to get a forecast of the FTSE 100 index call option price of tomorrow. If the option price forecast is greater than the market price of the option today, the call option is bought. If the option price forecast is less than the market option price today, the call option is sold. Note, however, that we apply a filtering strategy where each agent trades only when the price change is expected to exceed 2% of today's price,[12] i.e.

$$(\hat{C}_{i(t+1)} - C_{it})/C_{it} > 2\% \tag{7.23}$$

Clearly, the trading strategies of the agents (i.e. trading calls) are not delta-neutral. However, unlike Noh *et al.* (1994), who used delta-neutral straddles to test their agents solely for their volatility forecasts, the essence of our trading exercise is to speculate on the joint direction of price and volatility. Of course, in effect only the Bayesian agents (i.e. Agents 4 and 5), who construct their forecasts based on the joint predictive density of price and volatility, have the tool to do so. Agents 1, 2 and 3, who use *ad hoc* versions of Black–Scholes, might be viewed as making pure volatility bets with delta exposure. In effect they do so, but they do it consciously. It is just that their view on the direction of the underlying is that it will remain fixed at its current value. In any case, to avoid large delta-exposures we restrict the positions of the traders to last only a day. Thus all the traders are forced to close their position tomorrow. Hence for a long position the trader sells the option at tomorrow's settlement price; for a short position the trader buys the option at tomorrow's settlement price.

Each agent is given £100 to invest each time. When a call option is sold, we allow the agent to invest the proceeds plus £100 in a risk-free asset. Also, if the forecasting rule indicates that no trade should take place, the sum of £100 is invested in a risk-free asset. For simplicity, we assume that the rate on the risk-free asset is zero. Thus, the rate of return (RT) (per trade) on buying call options is computed as:

$$RT = \frac{100}{C_t}(C_{t+1} - C_t) \tag{7.24}$$

The rate of return (RT) on selling call options is computed as:

$$RT = \frac{100}{C_t}(-(C_{t+1} - C_t)) \tag{7.25}$$

The above rates, however, are without taking into account transaction costs. We need to incorporate this. Hence we assume that the transaction costs (per trade) incurred by the

[12] The cut-off value in the filtering rule can be chosen in relation to the transaction costs incurred by the agents. For example, in the following page we assume that transaction costs amount to 2% of the amount invested. Thus the cut-off of 2% implies that agents are willing to trade only if they think they will be able to recover at least the transaction costs.

agent amount to 2% of the amount invested, and so the net rate of return (NRT) from the trading of options is computed by:

$$NRT = RT - 100^{*}2\% = RT - 2 \tag{7.26}$$

Noh *et al.* (1994) assume that the transaction cost for trading a straddle (i.e. a call and a put option) is \$0.25 per straddle. Inspired from that, we also calculate an alternative NRT were we assume that the cost of trading a call option is £0.50:

$$NRT = RT - \frac{100}{C_t} * 0.50 \tag{7.27}$$

Discussions with option traders suggest that our transaction costs are not accurate, as they do not capture the huge spreads and extreme illiquidity that can occur in these markets. Thus our trading profits should be seen as a measure of economic worth and not necessarily as an attainable amount of money.

We can now compare the performance of the agents with different forecasting algorithms. In Table 7.3, we report the mean return of each agent per trade (or day).

Table 7.3 shows the daily rate of return from trading call options before and after transaction costs. It is clear that Agents 4 and 5 outperform the others, with average daily rates of return between 4.6% and 6.6% and 4.2% and 6.2% respectively, depending on the transaction costs incurred. It should be noted, though, that the profits of all agents are far from certain, since the corresponding standard deviations range from 19.6 to 29. However, t-ratios higher than 2 indicate that profits from the Bayesian forecasting

Table 7.3 Mean daily rate of return (or mean per trade rate of return) from trading FTSE 100 Index European call options

	Mean	SD	t-Ratio
Before transaction costs			
Agent 1: (historical volatility)	1.4%	29.0	0.46
Agent 2: (implied volatility)	2.3%	19.8	1.11
Agent 3: (today's implied volatility)	0.4%	25.5	0.15
Agent 4: (Bayesian)	6.6%	25.0	2.52
Agent 5: (Bayesian A-T-M)	6.2%	26.7	2.22
2% transaction costs			
Agent 1: (historical volatility)	−0.6%	29.0	0.20
Agent 2: (implied volatility)	0.3%	19.8	0.15
Agent 3: (today's implied volatility)	−1.6%	25.5	0.60
Agent 4: (Bayesian)	4.6%	25.0	1.76
Agent 5: (Bayesian A-T-M)	4.2%	26.7	1.51
£0.50 per call option transaction costs			
Agent 1: (historical volatility)	0.4%	28.2	0.14
Agent 2: (implied volatility)	1.7%	19.6	0.83
Agent 3: (today's implied volatility)	−0.4%	26.2	0.15
Agent 4: (Bayesian)	5.8%	24.2	2.29
Agent 5: (Bayesian A-T-M)	5.4%	26.0	1.98

Table 7.4 Cumulative returns from call option trading

	Before transaction costs	2% transaction costs	£0.50 per call option transaction costs
Agent 1	128.4%	−138.6%	39.9%
Agent 2	212.2%	38.2%	156.2%
Agent 3	38.8%	−123.2%	−33.7%
Agent 4	603.1%	417.1%	531.7%
Agent 5	567.5%	354.5%	490.7%

method are significantly greater than zero.[13] This argument is supported by Table 7.4, which shows the cumulative rate of return of the five agents.

Note that Table 7.4 is for comparative purposes among the different performances of the agents, and is not representative for assessing the cumulative return of each method over the 8-year period.

What is striking from Tables 7.3 and 7.4 is not the absolute performance of the agents, which (as already mentioned) may not necessarily represent real returns, but their comparative performance. Indeed, Agents 4 and 5 outperform comfortably the agents that use the standard routines. Noh *et al.* (1994) perform a similar study. They feed an asset's return series into a GARCH model to obtain a forecast of volatility to be plugged into the Black–Scholes model. They compare this method against forecasts obtained from implied volatility regressions that are also to be plugged into the Black–Scholes model. They assess the performance of these two volatility prediction models for S&P 500 index options over the April 1986 to December 1991 period. They find that the average daily rate of return from trading near-the-money straddles (before transaction costs) is 1.36% for the GARCH forecasting method and 0.44% for the implied volatility forecasting method. The standard deviations that they obtain are also quite large (in the range of 10–12). Although because of the large standard deviations we cannot really be sure, observe that what we find in Table 7.3 is not that dissimilar. Agent 1, who uses historical returns, delivers (before transaction costs) a mean return of 1.4%, and Agents 2 and 3, who use implied volatilities, deliver 2.3% and 0.4% respectively.

Finally, in Table 7.5 we report the percentage of times where the forecast of each agent resulted in a profit being made, a no-trade situation, or a loss being made.

Table 7.5 Percentage of times where the forecast of each agent resulted in a profit being made, a no-trade situation, or a loss being made

	Profit	No trade	Loss
Agent 1: (historical volatility – 30 days)	52.7%	4.4%	42.9%
Agent 2: (implied volatility – 30 days)	34.0%	37.4%	28.6%
Agent 3: (today's implied volatility)	30.8%	41.7%	27.5%
Agent 4: (Bayesian – implied volatility)	46.2%	34.1%	19.7%
Agent 5: (Bayesian – A-T-M implied volatility)	49.5%	24.2%	26.3%

[13]Since rates of return from call trading are assumed to be independent, the t-ratio is computed as a ratio of mean to standard deviation divided by the square root of the number of observations.

In this case as well, Agents 4 and 5 outperform the others. They make a profit 70% and 65% respectively of the times they *trade*. (The loss-making times being, of course, the remaining percentage.) Although Agent 1 in absolute terms makes a profit more times than every other agent (i.e. 52.7%), in real terms her profit rises to only 55% of the times she trades. Finally Agents 2 and 3 make a profit 54% and 53%, respectively, of the times they trade.

7.5 Conclusion

It has been suggested that both historical and implied volatilities convey information about future volatility. In this chapter we have developed a formal Bayesian framework to simultaneously exploit the information content of historical data as represented in moving averages of daily squared returns with the information content of options prices as represented in moving averages of reported implied volatilities. To this end, we have derived the posterior and predictive distributions of the Black–Scholes option price. We have used the FTSE 100 index European options market to compare our model's forecasting performance with standard models that use historical or implied volatility forecasts. All such benchmark forecasts are plugged into Black's (1976) model.

Our approach gives a modest outperformance relative to the usual *ad hoc* volatility schemes when measured in terms of MFE and RFE. However, when we assess our model in economic terms, i.e. in terms of the profit to a trading strategy based on our forecasts versus the other benchmark forecasts, we find quite substantial outperformance. We do not claim guaranteed excess risk adjusted returns for practitioners who might wish to follow our strategy. We recognize that option markets tend to have high and varying transaction costs and high illiquidity at unpredictable times, which makes real-time backtesting extremely difficult. Our results should be interpreted as an alternative measure of forecasting performance. With such an interpretation, our results indicate a clear superiority over the other methods.

We have not experimented a great deal with different weighting schemes for the prior and sample information, which, given our results, might deserve more attention. Clearly neither the particular weighting schemes nor the trading strategies considered in the chapter can be said to be optimal, and there is plenty of room for improvement. For example, a useful extension might be to attach weights according to the forecasting performance that each source of information has (i.e. estimate how well the implied and historical data explain volatility separately, and then simply combine the forecasts using a model weighting scheme). Likewise, we have only used the simplest of time-series models (i.e. moving average) to capture time variation in historical and implied volatilities. The sole reason for doing so was for higher theoretical consistency with our benchmark option-pricing model (i.e. the Black–Scholes). Although we did not pursue this point, we could have just as easily merged EWMA (exponentially weighted moving average) or GARCH volatility estimates with the implied volatility information. Given the success of GARCH when compared with other time-series models in capturing time-varying risk and the wealth of information about price risk contained in options, the combination of the two might have produced even better results. This would, however, lead us to the Duan (1995) GARCH option-pricing framework and to the Bauwens and Lubrano (2000) Bayesian GARCH approach. Finally, it would be very interesting and potentially more

fruitful to test the performance of the Bayesian framework in exercises of long(er)-run predictability of options prices as opposed to the 1 day ahead forecasts considered in this chapter. We leave all such analyses for future research.

Appendix: Analytic formulae for the densities required in deriving the posterior density of the Black–Scholes option price

Equation (7.12)

$$f(\mu, \sigma \backslash t, s', t', m) = \sqrt{\frac{2t}{\pi}} \frac{(t's'^2/2)^{t'/2}}{\Gamma(t'/2)} \frac{1}{\sigma^{t'+2}} \exp\left(-\frac{t's'^2 + t(\mu - m)^2}{2\sigma^2}\right)$$

Equation (7.13)

$$f(P_t, \mu, \sigma \backslash t, s', t', m) = \frac{1}{\pi P_t} \frac{(t's'^2/2)^{t'/2}}{\Gamma(t'/2)} \frac{1}{\sigma^{t'+3}} \times \exp\left\{-\frac{t's'^2}{2\sigma^2} - \frac{1}{2\sigma^2 t}\left(\left[\ln\left(\frac{P_t}{P_0}\right) + \frac{\sigma^2 t}{2} - \mu t\right]^2 + t^2(\mu - m)^2\right)\right\}$$

Equation (7.14)

$$f(P_t, \sigma \backslash t, s', t', m) = \frac{1}{\sqrt{\pi t} P_t} \frac{(t's'^2/2)^{t'/2}}{\Gamma(t'/2)} \frac{1}{\sigma^{t'+2}} \exp\left(-\frac{t's'^2}{2\sigma^2} - \frac{1}{4\sigma^2 t}\left[\ln\left(\frac{P_t}{P_0}\right) - \left(m - \frac{1}{2}\sigma^2\right) t\right]^2\right)$$

Equation (7.15)

$$f(P_t, \sigma \backslash s, t, s', t', m) = \frac{1}{\sqrt{\pi t} P_t} \frac{\left(\frac{t's'^2 + (t-1)s^2}{2}\right)^{(t'+t-1)/2}}{\Gamma\left(\frac{t'+t-1}{2}\right)} \frac{1}{\sigma^{t'+t+1}} \times \exp\left(-\frac{t's'^2 + (t-1)s^2}{2\sigma^2} - \frac{1}{4\sigma^2 t}\left[\ln\left(\frac{P_t}{P_0}\right) - \left(m - \frac{1}{2}\sigma^2\right) t\right]^2\right)$$

Equation (7.16)

$$f(s \backslash t, s', t', (m)) = \frac{2}{\mathrm{B}\left(\frac{t-1}{2}, \frac{t'}{2}\right)} \left(\frac{t's'^2}{2}\right)^{t'/2} \left(\frac{t-1}{2}\right)^{\frac{t-1}{2}} \frac{s^{t-2}}{\left(\frac{t's'^2 + (t-1)s^2}{2}\right)^{\frac{t+t'-1}{2}}}$$

Equation (7.17)

$$pdf(P_t, C_t \backslash s, t, s', t', m) = \frac{|J|}{\sqrt{\pi t} P_t \Gamma\left(\frac{t'+t-1}{2}\right)} \frac{\left(\frac{t's'^2+(t-1)s^2}{2}\right)^{\frac{t'+t-1}{2}}}{C_{BS}^{-1}(P_t, C_t)^{2\theta+\upsilon+2}} exp\left(-\frac{t's'^2+(t-1)s^2}{2C_{BS}^{-1}(P_t, C_t)^2}\right)$$
$$\times \exp\left(-\frac{1}{4C_{BS}^{-1}(P_t, C_t)^2 t}\left[\ln\left(\frac{P_t}{P_0}\right) - \left(m - \frac{1}{2}C_{BS}^{-1}(P_t, C_t)^2\right)t\right]^2\right)$$

Denominator of equation (7.18)

$$pdf(P_t \backslash s, t, s', t', m) = K_{\frac{t'+t}{2}}\left(\frac{1}{4}\sqrt{2\left(t's'^2+(t-1)s^2\right)t + \left(\ln\left(\frac{P_t}{P_0}\right) - mt\right)^2}\right)$$
$$\times \frac{\left(\frac{t's'^2+(t-1)s^2}{2}\right)^{(t'+t-1)/2}}{\sqrt{\pi t} P_t \Gamma\left(\frac{t'+t-1}{2}\right)} \left(\frac{t}{16}\right)^{\frac{t'+t}{2}}$$
$$\times \left[\frac{2(t's'^2+(t-1)s^2)t + \left(\ln\left(\frac{P_t}{P_0}\right) - mt\right)^2}{64}\right]^{-\frac{t'+t}{4}} \exp\left(-\frac{\ln(P_t/P_0) - mt}{4}\right)$$

where $K_{\frac{t'+t}{2}}\left(\frac{1}{4}\sqrt{2(t's'^2+(t-1)s^2)t + (\ln(\frac{P_t}{P_0}) - mt)^2}\right)$ is the modified Bessel function of the second kind of order $\frac{t'+t}{2}$.

References

Amin, K. and Jarrow, R. (1991). Pricing foreign currency options under stochastic interest rates. *Journal of International Money and Finance*, 10:310–329.

Baillie, R. and Bollerslev, T. (1992). Prediction in dynamic models with time-dependent conditional variances. *Journal of Econometrics*, 52:91–113.

Bauwens, L. and Lubrano, M. (2000). Bayesian option pricing using asymmetric GARCH models. Discussion Paper, CORE, Université catholique de Louvain.

Bauwens, L., Lubrano, M. and Richard, J.-F. (1999). *Bayesian Inference in Dynamic Econometric Models*. Oxford: Oxford University Press.

Black, F. (1976). The pricing of commodity contracts. *Journal of Financial Economics*, 3:167–179.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81:637–659.

Cox, J. and Ross, S. (1976). The valuation of options for alternative stochastic processes. *Journal of Financial Economics*, 3:145–166.

Darsinos, T. and Satchell, S. (2001). Bayesian analysis of the Black–Scholes option price. DAE (Department of Applied Economics) Working Paper No. 0102, University of Cambridge.

Day, R. and Lewis, C. (1992). Stock market volatility and the information content of stock index options. *Journal of Econometrics*, 52:267–287.
Derman, E. and Kani, I. (1998). Stochastic implied trees: arbitrage pricing with stochastic term and strike structure of volatility. *International Journal of Theoretical and Applied Finance*, 1(1):61–110.
Duan, J. (1995). The GARCH option pricing model. *Mathematical Finance*, 5(1):13–32.
Duan, J. and Zhang, H. (2001). Pricing Hang Seng index options around the Asian financial crisis – a GARCH approach. *Journal of Banking and Finance*, 25:1989–2014.
Dumas, B., Fleming, J. and Whaley, R. (1998). Implied volatility functions: empirical tests. *Journal of Finance*, 53:2059–2106.
Dupire, B. (1994). Pricing with a smile. *Risk*, 7:32–39.
Engle, R. and Mustafa, C. (1992). Implied ARCH models from options prices. *Journal of Econometrics*, 52:289–311.
Gemmill, G. and Saflekos, A. (2000). How useful are implied distributions? Evidence from stock-index options. *Journal of Derivatives*, 7(3):83–98.
Hafner, C. and Herwartz, H. (1999). Option pricing under linear autoregressive dynamics, heteroscedasticity and conditional leptokurtosis. Paper presented at ESEM99.
Hamilton, J. (1994). *Time Series Analysis*. Princeton, NJ: Princeton University Press.
Harvey, C. and Whaley, R. (1992). Market volatility prediction and the efficiency of the S&P 100 Index option market. *Journal of Financial Economics*, 31:43–73.
Heston, S. (1993). A closed form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6:327–343.
Heston, S. and Nandi, S. (2000). A closed-form GARCH option valuation model. *Review of Financial Studies*, 13:585–625.
Hobson, D. and Rogers, L. (1998). Complete models with stochastic volatility. *Mathematical Finance*, 8(1):27–48.
Hull, J. and White, A. (1987). The pricing of options on assets with stochastic volatilities. *Journal of Finance*, 42(2):281–300.
Karolyi, G. (1993). A Bayesian approach to modelling stock return volatility for option valuation. *Journal of Financial and Quantitative Analysis*, 28:579–594.
Latane, H. and Rendleman, R. (1976). Standard deviation of stock prices implied in option prices. *Journal of Finance*, 31:369–381.
Nelson, D. (1990). ARCH models as diffusion approximations. *Journal of Econometrics*, 45:7–38.
Noh, J., Engle, F. and Kane, A. (1994). Forecasting volatility and option prices of the S&P 500 index. *Journal of Derivatives*, 2:17–30.
Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Boston, MA: Harvard Business School.
Rebonato, R. (1999). *Volatility and Correlation: In the Pricing of Equity, FX and Interest-Rate Options*. Wiley Series in Financial Engineering. New York, NY: John Wiley & Sons.
Rubinstein, M. (1994). Implied binomial trees. *Journal of Finance*, 49:771–818.
Satchell, S. and Timmermann, A. (1993). Option pricing with GARCH and systematic consumption risk. Financial Economics Discussion Paper (FE/10), Birkbeck College, University of London.
Scott, L. (1987). Option pricing when the variance changes randomly: theory, estimation and an application. *Journal of Financial and Quantitative Analysis*, 22:419–438.
Stein, E. and Stein, C. (1991). Stock price distributions with stochastic volatility: an analytic approach. *Review of Financial Studies*, 4:727–752.
Wiggins, J. (1987). Option values under stochastic volatility: theory and empirical estimates. *Journal of Financial Economics*, 19:351–372.
Zellner, A. (1971). An Introduction to Bayesian Inference in Econometrics. New York, NY: Wiley.

8 Robust optimization for utilizing forecasted returns in institutional investment

Christos Koutsoyannis and Stephen Satchell

8.1 Introduction

The purpose of this chapter is to examine the usefulness of ideas of robust optimization. We shall look at potential applications in the area of financial forecasting. In keeping with the general theme of this book, we shall focus on expected returns forecasting, and how forecasting and robust optimization can be entered into jointly in a beneficial way. In the process of carrying out these tasks, we shall overview much of the current research in robust optimization; not least because it is an area of great commercial interest.

Robust optimization can have wide applications in finance, including in portfolio construction, option pricing and hedging, and risk management. Robust optimization deals with uncertainty. The robust optimization literature tries to address directly both parameter uncertainly (arising because of the imperfect data available to practitioners and researchers), and also model uncertainty (or the possibility of incorrect structure in the underlying model used). The purpose is to come up with more stable, relevant portfolios that are less prone to data and other errors. We will examine these concepts in more detail.

Note that there is an important distinction in this literature between risk and uncertainty. In simple terms, risk is when you can capture model variation by assigning priors and probabilities to parameters or structures; uncertainty is when you cannot assign probabilities. The distinction is profound, not least because of the experimental evidence of ambiguity (uncertainty) aversion, which is illustrated by Knight's and Ellsberg's Urns. These are stylized experiments, where investors have to choose an urn containing different mixes of coloured balls. We shall expand on these later.

In our review of robust optimization, we shall concentrate on potential uses to practitioners in building investment portfolios. We will consider extensions to standard quadratic utility problems, i.e. extensions to standard Markowitz mean-variance analysis. At the risk of tremendous oversimplification, there seem to be two very broad approaches. There is a mathematical formulation of robust optimization. This looks at more complex structures for an objective function or constraints. Solutions to such problems have become feasible because of advances in the mathematical literature, such as Interior Point solutions to second-order cone programming problems. This is the approach used by some practitioners. However, there is a second approach, much favoured by academics, which takes a maximin approach to expected utility problems, and is based upon what is

known in decision theory as ambiguity aversion. This is very important in understanding the intuition behind the robust techniques available.

We present a survey of both approaches to optimization. In particular, we look at the earlier analysis of Lutgens and Schotman (2004), which is an application of ambiguity aversion analysis to mean-variance problems. We also illustrate how to implement the different cases of what we would call multiple scenarios (neo-Bayesianism) versus multiple priors with associated prior probabilities (classical Bayesianism). We also present a case study with an implementation of a robust optimization problem.

The rest of the chapter is structured as follows. In Section 8.2, we review various robust procedures that might be of use to an investor. Section 8.3 is a case study, where we outline an example formulation for a robust optimization problem, based on forecast errors and extra quadratic constraints. In Section 8.4 we present some extensions to the theory and consider an application of robust ambiguity analysis to a stock selection model. Conclusions follow in Section 8.5.

8.2 Notions of robustness

There is a large literature on uncertainty and robust techniques. These occur in many fields, particularly mathematics, finance, economics and statistics, and with broad applications in investment, operations control and engineering. In this section we consider the application of robust concepts of interest to an institutional investor. Our principal interest here will be on forecasting returns, but, as will become clear to the reader, this will involve a wide range of robust concepts, all of which will have relevance to the forecasting problem.

It is apparent that robustness is a very broad subject and a complete survey would include thousands of references. We will concentrate on the areas we find relevant to forecasting returns and investment choices. We now present a list of the broad topics we shall consider.

Initially we shall review robust statistics and ranked forecasts, followed by a discussion of Bayesian adjustments, including shrinkage and Black–Litterman, as well as a discussion of risk aversion. Moving more to the optimization side of the problem, we look at the role of constraints, resampling as well as utility bounds. We end with a discussion of the mathematics of robust optimization, including second-order cone programming.

Our first notion of robustness is the theory of robust statistics. There are many different variants of this, and it is a large area in its own right. We will consider general robust estimators, and then concentrate on the work of Victora-Feser. Huber (1964) and Hampel (1974) both consider issues of robust estimation, and the effect of data errors on estimators. In fact, many robust estimators or robust versions of estimators have been developed to deal with data errors (for example, Least Trimmed Squares (Rousseeuw, 1985), Iteratively Reweighted Least-Squares (Holland and Welsch, 1977), Generalized Method of Moments estimation (Ronchetti and Trojani, 2001), *M*-estimators (Hampel *et al.*, 1986), *S*-estimators (Rousseeuw and Yohai, 1984), or τ-estimators (Yohai and Zamar, 1988)).

Victoria-Feser considers robust statistics in the framework of expected returns and robust portfolio selection. Victoria-Feser (2000) considers what happens to conventional

statistical techniques if the distribution of returns is contaminated by mixing it with some other distribution, typically a one-off extreme stock. So, for example, we may wish to examine what happens to return generating models by regarding the huge price movement of October 1987 as such a contamination. Perret-Gentil and Victoria-Feser (2003) expand on the above, and show that the use of robust estimates in portfolio selection problems result in more stable portfolios, mitigating estimation error and more importantly model uncertainty.

An alternative notion of robustness is based on using weaker information on the return forecasts. Thus we might consider only looking at rankings of stocks rather than actual point estimates of returns. This can be thought of as expressing robustness through incomplete information. The basic idea is that we would regard conventional return forecasts as so unreliable that we might only be willing to produce rankings. This is indeed what many funds do. Again this in itself would not exclude optimization. Indeed, risk could be measured conventionally. We would then need to combine ranked returns with conventional risk. This might involve adjusting the coefficient of risk aversion or mapping back forecasts to returns (see Satchell *et al.*, 2003; Satchell and Wright, 2005).

The approach outlined above has been criticized by Almgren and Chriss (2004, 2006) in an influential paper that is republished in this book (see Chapter 4). They have preferred an approach based on using cones. Each region of equal returns in n-dimensional space is defined as a cone, within which the true expected returns are believed to lie. A probability measure P is assigned defining the relative probability of the true return vector being in any location within the cone. These are linked to a constraint set within which the optimal portfolio should lie. For a single expected return vector, as in a standard Markowitz case, a portfolio is preferred to another one if it has a higher expected return. Given the (multiple) feasible expected return vectors in the cone, the authors deem a portfolio A to be superior to another B when the n-dimensional area where A's returns are higher is larger than the area where B's returns are larger, under the probability measure P above. Note that the authors find the centroid, or n-dimensional geometric centre-of-mass of the consistent set, to define perfectly this probability measure P. The process of finding an optimal portfolio is therefore reduced to a manipulation of this centroid. The intuition is to use minimal information on alpha to get a large feasible set of portfolios that might be dominating. This procedure is similar to some sophisticated manager performance systems that compare the manager's portfolio weights with all randomly generated weights that satisfy the same portfolio constraints. The comparison is then between the manager and the average portfolio, that is, the centroid.

There is a broad range of Bayesian adjustments, both by practitioners and in the academic literature. James and Stein (1961), Jobson *et al.* (1979) and Jorion (1986) have produced biased-shrinkage estimators applied to the mean. Note that these Bayesian adjustments are not necessarily restricted to forecasted returns, but can be applied to the other inputs to an optimizer as well. The idea is again exactly the same, where sample information is shrunk towards model-based *ex ante* information, again to deal with estimation uncertainty. For example, Ledoit and Wolf (2003) suggest shrinking the sample covariance matrix towards a single-index covariance matrix, and Ledoit and Wolf (2004) shrink it towards a constant correlation matrix to control for estimation error. The optimal shrinkage intensity is derived by minimizing a loss function based on the Frobenius norm of the matrix; see their paper for details.

With respect to improving forecasted returns, Black and Litterman (1990, 1992) proposed an intuitive way of combining model or equilibrium views with the views of the investor. Their suggestion was to start from equilibrium returns when optimizing (or by implied expected equilibrium returns as extracted by a reverse optimization method), and tilting towards the investor's views according to the degree of confidence of the investor on these views. The result should be a more stable portfolio. While the original paper was on asset allocation, the same methodology can be extended to an equities factor model. Lee (2000) suggests that the Black–Litterman model 'largely mitigates' the error-maximization issue of optimizers by spreading the errors across the vector of expected returns. Kandel and Stambaugh (1996), Pastor and Stambaugh (2000), Barberis (2000) and Pastor (2000) consider similar models of Bayesian updating. A number of chapters in this book discuss Black and Litterman's influential model.

Cavadini *et al.* (2001) concentrate on the importance of risk aversion on portfolio choice. Specifically, they consider what they call pseudo risk aversion corrections to explicitly address both estimation risk and model risk at the same time. In effect, they penalize risky allocations that are particularly exposed to both types of risk. They find substantial reductions in their loss function as a result, and superior results to addressing either estimation risk or model risk separately.

Practitioners routinely use constraints in the optimization to control for estimation error. In fact, Jagannathan and Ma (2003) demonstrate that even imposing wrong constraints can help. They show that imposing a non-negativity constraint to a portfolio optimization process is equivalent to shrinking extreme elements of the covariance matrix. In fact, the reduction of sampling error is often higher than the increase of specification error, resulting in less risky portfolios overall.

Michaud (1989, 1998) proposed a re-sampled frontier of optimal portfolios to deal with ambiguity aversion. He assumes that the true population is multivariate normal, where the mean and covariance matrix are given by a maximum likelihood estimator on historical data. Based on simulation over uncertain parameters, repeated mean-variance optimizations and an averaging of the optimal portfolios result in a more stable and robust solution, addressing the 'error-maximization' issue.[1] One can interpret this as a resolution of ambiguity aversion by treating each Monte Carlo resolution as a prior, and each prior as equally likely. It is not clear that a simple average of the optimal portfolios, proposed by Michaud, captures the distributional properties of the underlying uncertainty, nor that it is consistent with a maximum expected utility framework. Harvey *et al.* (2004) concentrate on the importance of higher moments and parameter uncertainty, pointing to Jensen's inequality problem with Michaud's re-sampling. This would imply that the re-sampled portfolio would converge in probability to the population expected value of the optimized portfolio, which will be typically biased in finite samples. For our purposes, it is not clear that, say, bequeathing some combined prior distribution to the re-sampled portfolios would lead to any theoretical or practical advantage over the robust techniques. Also, while intuitive in its conception, the computational overload of re-sampling might be a negative consideration for the practical investor.

Another robust procedure is to use techniques that give you lower bounds on your expected utility. Clearly, if this lower bound is valid for all scenarios, it will encompass

[1]The error-maximization issue refers to the estimation-error insensitivity of mean-variance optimizations, whereby for example excessive weights can be placed on unrealistically high estimates. Jobson and Korkie (1980) were the first to point to this bias.

the analysis that we discuss later. Such an approach is discussed in Anderson *et al.* (2000). More precisely, they use an exponential inequality that bounds the tail probability of final wealth (a Markov inequality). This procedure has been applied independently, to build portfolios with robust risk characteristics, by Chu (2004).

So far we have considered the role of uncertainty and uncertainty aversion when modelling preferences and resolutions based on treating the inputs or the outputs of a standard quadratic optimizer. However, there have recently been improvements in the mathematical and optimization literature itself.

Quadratic programming, while consistent with the Markowitz risk-return framework, can be restrictive with regards to real-life problems as it can only handle linear constraints. Non-linear optimization can only deal with a limited problem size, because the time taken to optimize increases exponentially with the number of assets. It can also lead to local as opposed to global optima. Rockafellar (1993) famously notes that in fact it is not the non-linearity *per se* that slows a solution down, but rather the non-convexity of the objective function or the constraints. This observation sparked a series of advancements. Nesterov and Nemirovsky (1994) developed an interior point method for non-linear convex optimization problems. This allows an investor to define a more general class of problems, with any number of quadratic constraints, which can still be solved efficiently (see for example Ben-Tal and Nemirovski (1998), or Lobo *et al.* (1997) for an implementation, including a link to sample code). In fact, primal-dual interior-point or barrier methods have been developed for various classes of problems. Other advances in optimization methodologies include integrating the maximin expected utility theory of Gilboa and Shmeidler (1989) mathematically (as opposed to via the objective function) into robust optimization methods. Lobo (2000) and Cornuejols and Tütüncü (2006) provide excellent reviews of the mathematical advances in robust optimization, with applications in finance.

The above discussion has wandered a little from our objectives of understanding how forecast error and optimization might interface. For our purposes, uncertainty about the return forecasts can be explicitly treated via more complex objective formulations. For example, extra quadratic constraints can be defined with regard to the first and second moments of the historic forecast error for each asset. Using the interior point methods above, with the extra quadratic constraints, would control for assets where forecasts have historically been wrong or volatile.

As expected, there have been numerous applications of the above optimization advances in the financial literature. For example, Goldfarb and Iyengar (2003) show that their problem formulation, which includes 'uncertainty structures' for dealing with uncertainty in the estimation of the model parameters, is in fact a conic optimization problem. Second-order cone programming (SOCP) deals with linear objective functions and linear and SOC constraints.

Costa and Paiva (2002) also consider a portfolio optimization problem under parameter uncertainty (both for the return forecasts and the covariance matrix). They show it is possible to consider classes of problems that have an optimal solution which will stay unchanged as we vary other aspects of the problem. In fact, they show their problem formulation can be reduced to a linear matrix inequalities optimization problem. However, these linear matrix inequalities that the unknown parameters satisfy and other aspects of the problem do not seem to have any economic meaning – at least none is provided by the authors.

8.3 Case study: an implementation of robustness via forecast errors and quadratic constraints

This section sets up a robust portfolio optimization problem, where the uncertainty about return forecasts is addressed via extra constraints on the first two moments of Forecast Errors (FE). As has become obvious by now, there are very broad classes of problems that deal with parameter uncertainty or model uncertainty in different ways. The set-up we will build up in this section should be taken as a case study, and is meant to introduce the reader to a 'robust' way of thinking, as well as 'robust' techniques like second-order cone programming (SOCP).

Our first task is to differentiate between the 'informative' and 'noisy' data that go into the process. Controlling FE exposure and variance improves portfolio stability more than traditional methods because it directly addresses and limits the main cause of instability; the instability in returns expectations, limiting the effect of the 'noisy' and allowing the 'informative' to add value to the process.

The standard mean-variance portfolio optimization or utility-maximization problem, where the investor looks for those portfolio weights w that maximize the risk-adjusted expected return of the portfolio, given a degree of risk aversion λ and returns expectations α, can be represented as:

$$\max U(w, \lambda) = \alpha^T (w - b) - \frac{\lambda}{2}(w - b)^T C(w - b) \tag{8.1}$$

where $w =$ the vector of asset weights in the portfolio, $\lambda =$ the degree of risk aversion, $\alpha =$ the vector of expected returns, $b =$ the vector of asset weights in the benchmark, and $C =$ the covariance matrix of asset returns.

Note that the scaling coefficient λ, being in the units of 'marginal expected return with respect to variance', can be difficult to evaluate with confidence. In practice, users of optimization methods that require an objective function of the form in equation (8.1) tend to make an initial guess at a value for λ, and subsequently adjust the value to achieve the tracking error required. Many products automate this iterative procedure. Note that it might not be known from a mathematical perspective whether a solution is feasible or not, given other constraints levied on the strategy, leading to the failure of this iterative procedure.

The maximization problem in equation (8.1) above can be reformulated to have a risk constraint rather than a risk term in the objective function. Using a risk constraint in this way enables the user to put a straightforward constraint on tracking error, and know prior to execution whether or not the problem is feasible:

$$\max U(w, \lambda) = a^{\mathrm{T}} (w - b) \tag{8.2}$$

subject to the risk constraint

$$\sigma_P^2 = (w - b)^T C(w - b) < TE \tag{8.3}$$

Regardless of whether risk is a term in the objective function, or applied as a constraint, both equations (8.1) and (8.2) suffer from robustness issues. As there is uncertainty

associated with the vector of expected returns α, the output portfolio composition w can be disproportionately unstable. In practical situations, a portfolio manager might, for example, implement constraints as discussed in Section 8.2 above (see Jagannathan and Ma, 2003).

As a potential robust solution to this issue, we propose accounting directly for the first two moments of FE. Forecast error (FE), or the difference between the *ex ante* expectation of return and the subsequent realized return, is easily measured and is independent of the forecasting methodology. By setting non-linear boundaries on FE mean and variance, the process discounts the impact of historically unstable or erroneous alphas, and places more emphasis on assets that are consistently forecasted well (where FE is small and consistent). It should therefore avoid excess turnover due to noise, and result in more stable portfolios, which are less prone to data errors or misspecification (parameter or model uncertainty). We feel that it may offer a better approach than other variations on standard mean-variance optimization, such as re-sampling and bootstrapping.

In the context of this robustness adjustment, our optimization problem then becomes to maximize expected utility as captured by portfolio returns. Subject to (1) a tracking error constraint as before, (2) a mean FE constraint and (3) a FE variance constraint, as below. Note that, as well as the FE variance constraint being by nature quadratic, in this example the mean FE constraint is also quadratic in order. As such, as the forecast error increases in magnitude, we penalize stocks increasingly aggressively.

$$\max U(w, \lambda) = a^T(w - b) \tag{8.4}$$

subject to

$$(w - b)^T C(w - b) < TE \tag{8.5}$$

$$(w - b)^T \mu_{FE}(w - b) < M^2 \tag{8.6}$$

and

$$(w - b)^T C_{FE}(w - b) < S^2 \tag{8.7}$$

where μ_{FE} = a diagonal matrix of squared mean forecast errors, M = the maximum tolerable mean forecast error, C_{FE} = the covariance matrix of forecast errors, and S = the maximum tolerable forecast error variance.

Remember that the recent advances in the mathematical optimization literature discussed in Section 8.2, and especially interior point methods, allow such problems with multiple quadratic constraints (like equations (8.5), (8.6) and (8.7)) to be solved efficiently (see, for example, Lobo *et al.*, 1997 for an implementation).

Note that both FE constraints can be expressed in terms relative to the initial portfolio position. An alternate way of thinking of these constraints then is as 'smart turnover' constraints. They both limit the optimizer's ability to deviate from the initial position, except for the case where there is stable, historically accurate alpha. This might draw an explicit link between the concepts of robustness and turnover:

$$(w - w_i)^T \mu_{FE}(w - w_i) < M^2 \tag{8.8}$$

and

$$(w - w_i)^T C_{FE}(w - w_i) < S^2 \tag{8.9}$$

where w_i = initial portfolio weights.

It is interesting to note that, whichever way the FE constraints are expressed, forecasted returns are adjusted for any source of uncertainty. For example, consistent FEs can arise because of incorrect adjustments for differences in accounting standards across regions, or because forecasting factors have sector or style biases. Robust optimization methods, and specifically the formulation discussed in this section, are independent of the source of the bias.

8.4 Extensions to the theory

There is an academic approach to robustness that equates robustness with ambiguity aversion. The ideas are of great interest, and we present them below. The original contribution in this field is due to Ellsberg (1961) who presented a fascinating decision-theoretic paradox, which we will describe next.

To quote Anderson *et al.* (2000: 7):

> Ellsberg (1961) created an example that challenged the Bayesian–Savage model of decision-making. Ellsberg (1961) considered a choice between bets on two urns. In Urn A, it is known in advance that there are fifty red balls and fifty black balls. In Urn B, the fractions of red and black balls are not known in advance. A ball will be drawn randomly from each urn. At no cost a decision maker is permitted to guess the colour of the ball drawn from one and only one of the urns. If the decision maker guesses the colour drawn from the chosen urn, he receives a positive payoff. The Bayesian–Savage model predicts either indifference between urns or a preference for Urn B, depending on the prior probability assignment. But Ellsberg (1961) argued that a preference for Urn A is reasonable. To Ellsberg, there is an important distinction between urns for which you are informed about probabilities vis-à-vis ones for which you are not... The ambiguity about the assignment of priors is a way to make Knight's (1921) notion of uncertainty operational. An aversion to such uncertainty can rationalise a preference for Urn A.

Just to clarify why it is 'rational' to pick Urn B, note that, assuming Urn B has anything other than an equal number of red and black balls, it will lead to better odds than Urn A, where we know in advance that our chance of winning is exactly 50 per cent. Urn B can never be worse than that, yet we choose Urn A because, psychologically, it is better the devil you know.

This notion of ambiguity aversion was captured by Gilboa and Schmeidler (1989). Gilboa and Schmeidler (1989: 142, paras 1–2) present an ingenious argument relating ambiguity aversion to uncertainty in the presence of multiple priors. Considering the worst case for Urn B, you may bet on red when there are zero red balls in the urn. Likewise, you may bet on black when there are zero black balls in the urn. Such worst-case scenarios will have a payout of zero. This is obviously worse than the $50 you would expect to get from Urn A if the prize is $100.

This resolution is a situation where uncertainty is dealt with by multiple priors and some version of maximin expected utility used to determine one's final decision. The more

ambiguity-averse the investor, the more priors he or she would entertain and, generically, the lower the expected utility the investor would end up with. This is an exciting idea, because it gives a theoretical structure to the well-known practice of scenario-analysis. In the context of mean-variance analysis, this would mean that we would consider the minimum expected utility function taken over a number of scenarios, such as high/low risk and high/low return, for example. In this case, the scenario of low return, high risk would always give the lowest expected utility. But we can easily imagine examples where, in moving over different portfolio combinations, the investor moves from one regime to the other. Once we have determined the minimum of these regimes, the optimal portfolio is calculated by choosing those portfolio weights that maximize utility over this minimum path. The essential difference between the above and conventional Bayesian analysis is that we do not assign probabilities to the scenarios.

To move the discussion from mathematics to institutional investment, consider the decision by a pension fund trustee to invest in a hedge fund. This is clearly fraught with ambiguity aversion, as we can imagine many scenarios in the mind of the trustee. We can also see that regulated, exchange-traded assets will have, in the eyes of the investor, much less ambiguity aversion. Likewise, operational risk is an ambiguity-aversion situation. You may be able to compute most of the scenarios, but you cannot assign prior probabilities to their likely occurrence, either through lack of data or through failure to clarify what the appropriate states of the world are.

A discussion of mean-variance analysis, and the explicit form of the min utility function that arises in the two scenarios, one risky asset case, is presented in the appendix of Lutgens and Schotman (2004).

To illustrate this further, consider the case of two scenarios and one risky asset (see Figure 8.1). The graphs show expected utility as we vary the weight w of the risky asset in the portfolio. Instead of choosing portfolio AA or BB, we prefer portfolio CC since this is the maximum of the minimum function. Notice that this is not the same as the worst-case scenario, which would be BB, since BB has lower expected utility than AA. Notice also that the chosen portfolio is not necessarily terribly conservative, in our example, we hold more of the risky asset, e.g. equity, than AA but less than BB.

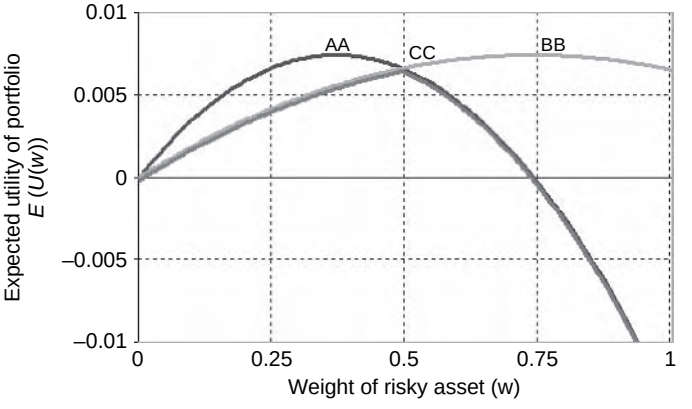


Figure 8.1 Maximin expected utility.

Up until now we have framed robustness solely in terms of ambiguity aversion, but there are numerous other approaches that have been considered, including in a portfolio construction framework.

Lutgens and Schotman (2004) consider an analysis based on linear factor models and use earlier work by Pastor and Stambaugh (2000). Their model essentially assigns priors to the covariance matrices of the factors and of the errors. They treat the factor exposures and factor returns as known. Other work in this area focuses on unconditional distributions but does not look at diversity of views about factor returns.

We shall consider a number of experts with differing views about implied factor returns, and ask how this would impact our optimal portfolio construction. If the experts agreed about all other aspects of the model, then we would have diversity of opinion on stock expected returns but nothing else.

Following the notation of Lutgens and Schotman (2004), we consider a k-factor model for returns as:

$$y_t = \alpha_t i + \beta_t V_t + u_t \tag{8.10}$$

$$E(u_t u_t') = D$$

where y_t and u_t are $(N \times 1)$ vectors, i is a $(N \times 1)$ vector of ones, β_t is an $(N \times k)$ matrix of (known) exposures, whilst V_t is the $(k \times 1)$ vector of (unknown) implied factor returns. The matrix D is a diagonal $(N \times N)$ covariance matrix and α_t is an unknown scalar. For the purpose of stock selection, α_t, being the same for all stocks, is unimportant, and to simplify matters we shall just consider:

$$y_t = \beta_t V_t + u_t \tag{8.11}$$

The weighted least squares estimator $\hat{V}_t$ is given by

$$\hat{V}_t = (\beta_t' \hat{D}^{-1} \beta_t)^{-1} (\beta_t' \hat{D}^{-1} y_t)$$

where $\hat{D}^{-1}$ is estimated by using residuals of previous periods to get estimates of idiosyncratic variance; it is recommended that rolling windows be used for this purpose. We write the 'true' value of V_t V_t^p. Furthermore, $\hat{V}_t \sim (V_t^p, (\beta_t' \hat{D}^{-1} \beta_t)^{-1})$. For the case of multiple priors, we can conceive of each expert j proposing $V_t^j \sim N(V_0^j, \Omega^j)$. In a Bayesian framework, we would assign probability π_j to the expert, based on past performance or pedigree or introspection.

The overall prior distribution is then a mixture of normals:

$$pdf(V_t^p) = \sum \pi_j N(V_o^j, \Omega^j)$$

If we combine this with the sample distribution of $\hat{V}_t$, we see that:

$$pdf(\hat{V}_t, V_t^p) = pdf(\hat{V}_t | V_t^p) pdf(V_t^p) = \sum \pi_j N(\mu_j, \theta_j)$$

where

$$\mu_j = ((\beta_t' \hat{D}^{-1} \beta_t) + \Omega_j^{-1})^{-1} [(\beta_t' \hat{D}^{-1} \beta_t)^{-1} V_t + \Omega_j^{-1} V_0^j] \tag{8.12}$$

It can be shown that the mean of a mixture of normals is equal to $\Sigma\pi_j\theta_j$ and the covariance matrix is equal to $\Sigma\pi_j\theta_j + \Sigma\pi_j\mu_j\mu_j' - (\Sigma\pi_j\mu_j)(\Sigma\pi_j\mu_j)'$ (see for example Satchell and Scowcroft (2000), reprinted in this book as Chapter 3).

In the case that we have a simple prior, we recapture the Black–Litterman model (1990, 1992; this has been briefly discussed above). In the case of ambiguity aversion, we would consider each separate distribution for $V_t \sim N(\mu_j, \theta_j)$, and calculate expected utility, as discussed before. Specifically, our measure of expected stock return would be $\beta_t\mu_j$, where μ_j is given by equation (8.12), and our measure of covariance $\beta_t\Psi\beta_t' + D$, that is we allow experts to agree on risk but differ on alpha. Based on these different alphas, we would compute $\mu_j(w) = w'\beta_t\mu_j - \lambda w'(\beta_t\Psi\beta_t' + D)w$. Note that Ψ is an estimate of the factor covariance matrix. We then compute $U(w) = \min_j(U_j(w))$ and optimize. The solution to this mean-variance problem will be like Theorem 1 in Lutgens and Schotman (2004). That is, our optimal portfolio will be of the form:

$$w^* = (\beta_t\Psi\beta_t' + D)^{-1}(\sum \lambda_j\beta_r\mu_j) \tag{8.13}$$

where the λ_j's are Lagrange multipliers essentially determining which of the experts' opinions is the most pessimistic for the particular combination of weights. These may be very difficult to compute numerically.

There is no agreed way of updating non-unique priors, although there is a huge literature in decision theory (see Gilboa and Schmeidler (1993) for some results and references). Our procedure of including all scenarios (priors) in the update can be seen as extreme in that we do not use sample information to gain information about the possible probabilities of priors. Gilboa and Schmeidler (1993) refer to work by Fagin and Halpern (1991) that follows the procedure we have adopted.

8.5 Conclusion

This chapter has provided an overview of robust optimization from the perspective of the practitioner and the academic. We note that practitioner interest focuses on mean-variance analysis using quadratic inequality constraints, whilst the academic literature is more concerned with issues about ambiguity aversion. Both are tremendously interesting, but the latter seems hard to implement for large portfolios with high levels of ambiguity. We illustrate how to implement the former using a forecasting problem based on constraining the error in our historic forecasts.

Acknowledgement

The authors would like to thank Bita Risk for useful discussions on robust optimization.

References

Almgren, R. and Chriss, N. (2004). Portfolio optimization without forecasts. Working Paper.

Almgren, R. and Chriss, N. (2006). Optimal portfolios from ordering information. *Journal of Risk*, 9(1), reproduced in this book as Chapter 4.

Anderson, E., Hansen, L. and Sargent, T. (2000). Robustness, detection, and the price of risk. Working Paper, Stanford University.

Barberis, N. (2000). Investing for the long run when returns are predictable. *Journal of Finance*, 55:225–264.

Ben-Tal, A. and Nemirovski, A. (1998). Robust convex optimization. *Mathematics of Operations Research*, 23(4).

Black, F. and Litterman, R. (1990). Asset allocation: combining investor views with market equilibrium. Technical Report, Goldman, Sachs, Fixed Income Research.

Black, F. and Litterman, R. (1992). Global portfolio optimization. *Financial Analysts Journal*, 48:28–43.

Cavadini, F., Sbuelz, A. and Trojani, F. (2001). A simplified way of incorporating model risk, estimation risk and robustness in mean variance portfolio management. Working Paper.

Chu, B. M. (2004). Using the large deviation theory to estimate a symmetric financial risk. PhD Dissertation, Birkbeck College, University of London.

Cornuejols, G. and Tütüncü, R. (2006). *Optimization Methods in Finance*. Cambridge: Cambridge University Press. Forthcoming.

Costa, O. and Paiva, A. (2002). Robust portfolio selection using linear-matrix inequalities. *Journal of Economic Dynamics and Control*, 26(6):889–909.

Ellsberg, D. (1961). Risk, ambiguity and the savage axioms. *Quarterly Journal of Economics*, 75:643–669.

Fagin, R. and Halpern, J. Y. (1991). A new approach to updating beliefs. In: P. P. Bonissone, M. Henrion, L. N. Kanal and T. Lemmer (eds), *Uncertainty in Artificial Intelligence*, Vol. VI. Amsterdam: Elsevier, pp. 347–374.

Gilboa, I. and Schmeidler, D. (1989). Maximin expected utility with a non-unique prior. *Journal of Mathematical Economics*, 18:141–153.

Gilboa, I. and Schmeidler, D. (1993). Updating ambiguous beliefs. *Journal of Economic Theory*, 59:33–49.

Goldfarb, D. and Iyengar, G. (2003). Robust portfolio selection problems. *Mathematics of Operations Research*, 28(1):1–38.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69:383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Function*. New York, NY: John Wiley.

Harvey, C., Liechty, J. C., Liechty, M. W. and Muller, P. (2004). Portfolio selection with higher moments. Working Paper.

Holland, P. W. and Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics: Theory and Methods*, A6:813–827.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101.

Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios: a role for portfolio weight constraints. Working Paper.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:311–319.

Jobson, J. D. and Korkie, B. (1980). Estimation for Markowitz efficient portfolios. *Journal of the American Statistical Association*, 75(371):544–555.

Jobson, J. D., Korkie, B. and Ratti, V. (1979). Improved estimation for Markowitz portfolios using James-Stein type estimators. *Proceedings of the American Statistical Association, Business and Economics Statistics Section*. Washington, DC: American Statistical Association.

Jorion P. (1986). Bayes–Stein estimation for portfolio analysis. *Journal of Financial and Quantitative Analysis*, 21(3):279–292.

Kandel, S. and Stambaugh, R. F. (1996). On the predictability of stock returns: an asset-allocation perspective. *Journal of Finance*, 51:385–424.

Knight, F. (1921). Risk, uncertainty, and profit. PhD thesis. Boston: Houghton Mifflin.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621.

Ledoit, O. and Wolf, M. (2004). Honey, I shrunk the sample covariance matrix. *Journal of Portfolio Management*: 31(1).

Lee, W. (2000). *Advanced Theory and Methodology of Tactical Asset Allocation*. New York, NY: John Wiley.

Lobo, M. S. (2000). Robust and convex optimization with applications in finance. PhD thesis, Stanford University.

Lobo, M. S., Vandenberghe, L., Boyd, S. and Lebret, H. (1997). Second order cone programming. Working Paper.

Lutgens, F. (2004). Robust portfolio optimization. PhD Thesis, University of Maastricht.

Lutgens, F. and Schotman, P. (2004). Robust portfolio optimisation with multiple experts. Inquire Europe Report.

Michaud, R. (1989). The Markowitz optimization enigma: is 'optimized' optimal? *Financial Analyst Journal*, 45(1):31–42.

Michaud, R. (1998). *Efficient Asset Management: A Practical Guide to Stock Portfolio Management and Asset Allocation*. Cambridge, MA: Harvard Business School Press.

Nesterov, Y. and Nemirovsky, A. (1994). Interior point polynomial methods in convex programming. *Studies in Applied Mathematics*, 13.

Pastor, L. (2000). Portfolio selection and asset pricing models. *Journal of Finance*, 55(1):179–223.

Pastor, L. and Stambaugh, F. (2000). Comparing asset pricing models: an investment perspective. *Journal of Financial Economics*, 56:335–381.

Perret-Gentil, C. and Victoria-Feser, M.P. (2003). Robust mean-variance portfolio selection. *Cahiers du Département d' Econométrie*, 2003.02, Université de Genève.

Rockafellar, R. T. (1993). Lagrange multipliers and optimality. *SIAM Review*, 35:183–283.

Ronchetti, E. and Trojani, F. (2001). Robust inference with GMM estimators. *Journal of Econometrics*, 101(1):37–69.

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In: W. Grossmann, G. Pflug, I. Vincze and W. Wertz (eds), *Mathematical Statistics and Applications*. Dordrecht: Reidel Publishing Co., pp. 283–297.

Rousseeuw, P. and Yohai, V. (1984). Robust regression by means of S-estimators. In: J Franke, W. Härdle and D. Martin (eds), *Robust and Nonlinear Time Series Analysis*, Lecture Notes in Statistics 26. Berlin: Springer Verlag, pp. 256–272.

Satchell, S. E. and Scowcroft, A. (2000). A demystification of the Black–Litterman model. *Journal of Asset Management*, 11(2):138–150.

Satchell, S. E. and Wright, S. M. (2005). Robust cross-sectional factor modelling approach to equity forecast construction. *Economic and Financial Modelling*, 12(4):153–182.

Satchell, S. E., Wright, S.M. and Huang, S. (2003). Assessing the merits of rank-based optimisation for portfolio construction. In: S. Satchell and A Scowcroft (eds), *New Advances in Portfolio Construction and Implementation*. London: Butterworth-Heinemann.

Victoria-Feser, M. P. (2000). Robust portfolio selection. *Cahiers de recherche HEC*, 2000.14, Université de Genève.

Yohai, V. J. and Zamar, R. H. (1988). High breakdown point estimates of regression by means of the minimization of an efficient scale. *Journal of American Statistical Association*, 83:406–413.

9 Cross-sectional stock returns in the UK market: the role of liquidity risk

Soosung Hwang and Chensheng Lu

Abstract

The relationship between liquidity and stock returns has been investigated extensively in recent years. Using UK data, we show that there is a sizeable difference in the cross-sectional returns between liquid and illiquid assets. Liquidity together with book-to-market equity explains cross-sectional returns. Furthermore, the well-documented value premium can be explained by a liquidity-augmented CAPM, and this result is robust in the presence of distress factors and a battery of macroeconomic variables. This suggests that liquidity could be a useful component in forecasting expected returns.

9.1 Introduction

Liquidity in the financial markets has been one of the critical issues in both practice and academia. Since the 1980s, a number of episodes of financial market distress have underscored the importance of the smooth functioning of markets for the stability of the financial system. At the heart of these episodes was a sudden and drastic reduction in market liquidity, characterized by disorderly adjustments in asset prices, a sharp increase in the costs of executing transactions, and so forth. The well-known 1998 episode involving Long-Term Capital Management (LTCM) is a representative example, and has prompted investors to care more about their liquidity risk when making portfolio decisions.

In this chapter, we investigate the role of liquidity risk in explaining cross-sectional stock returns. In particular, we examine the link between liquidity and the well-documented value premium. Fama and French (1992) point out that liquidity, though important, does not need to be specifically measured and accounted for, as it is subsumed by the combination of size and book-to-market factors. It is generally accepted that illiquid stocks tend to be small and that people would not be surprised to see the high correlation of size and liquidity. However, Brennan and Subrahmanyam (1996) show that there is a statistically significant positive relationship between expected stock returns and illiquidity, even after taking Fama–French risk factors into account. Additionally, Chordia *et al.* (2001) prove that liquidity does need to be accounted for individually, even after controlling for size, book-to-market and momentum. It is these stylized facts that link liquidity to forecasting expected returns.

Liquidity is a broad and elusive concept, which is not directly observed. Many liquidity proxies have been proposed, such as bid–ask spread, trading volume, or a combination of return and volume (see Section 9.2 for detailed descriptions). Among these liquidity measures, a few studies use trading volume as the proxy for the aggregate demand of liquidity traders (see Campbell *et al.*, 1993), which suggests there could be some link between liquidity and other factors. Lee and Swaminathan (2000) demonstrate that low (high) volume stocks display many characteristics commonly associated with value (growth) stocks. Therefore, the return spread between value and growth could contain the difference of liquidity risk inherited by them.

Since Fama and French (1992, 1993), many researchers have documented the existence of a value premium – i.e. the excess return of value stocks (high book-to-market) over growth stock (low book-to-market). Fama and French (1998) even find international evidence of this value anomaly. There are a growing number of studies that attempt to explain this value anomaly using different theories.[1] None of these, however, can successfully accounted for this value spread. Although Lee and Swaminathan (2000) document the empirical connection between trading volume and value/growth, they do not investigate the interaction between value/growth and liquidity. We formally test the relationship between value anomaly and liquidity risk in this study.

Our contribution is two-fold: first, we demonstrate that in the UK market there is a significant liquidity premium which cannot be explained by the CAPM, Fama and French three-factor model, or Fama and French with a momentum factor model.[2] Secondly, we provide evidence that liquidity explains the value premium. The value anomaly can be explained by a liquidity-augmented CAPM, which offers important implications for the link between value/growth and liquidity. Furthermore, the evidence of liquidity in explaining the value premium is not subsumed by the distress factor proposed by Agarwal and Taffler (2005) and a number of macroeconomic variables. The results are not consistent with those of Fama and French (1995, 1996) and Saretto (2004), who suggest the excess return of value over growth stocks is due to the distress risk inherited with them.

This chapter proceeds as follows. The next section describes the development of hypotheses and research designs. Section 9.3 shows the empirical results. The last section offers concluding remarks and future research directions.

[1] Zhang (2005) uses rational expectation theory in a neoclassical framework to explain this value anomaly. He finds that value is riskier than growth in poor market conditions when the price of risk is high and high book-to-market signals persistent low profitability. Petkova and Zhang (2005) find time-varying risk goes in the right direction in explaining value premium; however the beta-premium covariance in their study is still too small to explain the observed magnitude of the value anomaly. Other studies state that value spread is a premium for distress using a behavioural theory. These argue that this value anomaly is real but irrational, which is the result of investor's overreaction that leads to under pricing of value (distress) stocks and over-pricing of growth stocks.

[2] Overall, there is a considerable amount of literature about liquidity and asset pricing, but most research is done on the US market, with only a few investigations having been done on the UK market. As pointed out by Dimson *et al.* (2003a), 'It would be dangerous for investors to extrapolate into future from the US experience. We need to also look outside of the United States.' Thus the UK data is adopted in this research, which can answer the crucial question in asset pricing of 'whether the results obtained for the US stock markets can be generalized to markets in other countries'.

9.2 Hypotheses and calculating factors

The first part of this section explains the methods by which liquidity is employed to explain cross-sectional stock returns. The constructions of the liquidity measure and factor are then presented in the next subsection. We also explain how we construct size and value/growth factors, which may not necessarily be the same as the method used in the US market.

9.2.1 Liquidity effects and cross-sectional stock returns

Since Fama and French (1992, 1993), many empirical papers have documented that average stock returns are related to firm characteristics, such as size and book-to-market. These return patterns are apparently not explained by the CAPM, and are thus called anomalies. The value premium, i.e. the excess return of value stocks (high book-to-market) over growth stock (low book-to-market), has been extensively researched in the literature, with Fama and French (1998, 2006) providing additional international evidence for this value anomaly.

There has been a surge in studies that attempt to explain this value anomaly using different theories. Fama and French (1996) empirically demonstrate that, except for the short-term momentum, these anomalies largely disappear in the Fama–French three factor model. Ang and Chen (2005) show that the value premium can be explained by a conditional CAPM. Fama and French (2006), however, argue that Ang and Chen's (2005) evidence is specific to the period 1926–1963. Other studies, such as those of Zhang (2005) and Petkova and Zhang (2005), use different theories and methods to explain the value premium; however, their results show that the observed value premium is still too large to be explained. Overall, none of the research has successfully accounted for this value anomaly.

In this chapter, UK data are used so that our study can be treated as an out-of-sample investigation of the value premium. Similar to Ang and Chen (2005), the dependent variables in this study are cross-sectional stock return differences related to size and book-to-market. We form decile, quintile and thirtieth/seventieth percentile-breakpoint portfolios with regard to stock's characteristics such as size and book-to-market. Taking ten decile portfolios (Pi, t), for example, at the end of June in year t, ten size-decile portfolios are formed on stocks' ranked market value. Similarly, at the end of December year t based on the stock's book-to-market value, ten value-decile portfolios are formed. Then the dependent variable is a hedge portfolio that takes a long position in a small (high book-to-market) portfolio and a short position in a big (low book-to-market) portfolio ($P_{L,t} - P_{S,t}$), where $P_{L,t}$ stands for the long position of this portfolio, which consists of either small or high book-to-market stock groups; and $P_{S,t}$ refers to the short position of this portfolio, i.e. either big or low book-to-market portfolios. We denote the hedged portfolios according to the different breakpoints as S–B_d (H–M_d), S–B_q (H–M_q), and S–B_p (H–M_p), which corresponds to the decile, quintile and thirtieth/seventieth percentile-breakpoints respectively.

We test the CAPM upon ($P_{L,t} - P_{S,t}$) to see if there is any unexplained systematic risk – in other words, significant alphas. If the intercept (alpha) is significant, it suggests either that there is a failure of the CAPM or that there is an anomaly which cannot be explained by the CAPM.

Size is commonly referred to as one type of liquidity proxy, as investors would not expect the same level of liquidity between large and small stocks; thus, the return difference between small and big could be the result of the different liquidity risks associated with each of them. Campbell *et al.* (1993) argue that trading volume proxies for the aggregate demand of liquidity traders. Lee and Swaminathan (2000) demonstrate that low (high) volume stocks display many characteristics commonly associated with value (growth) stocks. Therefore, the return spread between value and growth should contain the differences in the liquidity risk inherited by them. In order to test these two hypotheses, we test the liquidity effects over the hedge portfolio, i.e. the following liquidity-augmented CAPM is estimated:

$$P_{L,t} - P_{S,t} = \alpha_i + \beta_{i,1}(Rm_t - R_f) + \beta_{i,2} LIQ_t + \varepsilon_{i,t} \tag{9.1}$$

The factor sensitivity for liquidity ($\beta_{i,2}$) should be significant in the above cross-sectional regression if liquidity effects are present.

If liquidity is one of the missing factors, it should be able to explain return anomalies to some extent. Following Pastor and Stambaugh (2003) and Ang and Chen (2005), the intercepts (alphas) of different portfolios strategies (such as, small minus big, high minus low) should not be significantly different from zero if liquidity is included.

9.2.2 Calculating factors in the UK market

Liquidity measures and liquidity factor

Like volatility, liquidity is not directly observed, and many different liquidity measures have been proposed for different purposes. Previous research, such as that by Amihud and Mendelson (1986), Chordia *et al.* (2000) and Hasbrouck and Seppi (2001), has focused on the bid–ask spread as a measure of illiquidity; however, as highlighted by Brennan and Subrahmanyam (1996), bid–ask spread is a noisy measure of illiquidity because many large trades appear outside the spread and many small trades occur within the spread. Brennan and Subrahmanyam (1996) proxy stock illiquidity using price impact, which is measured as the price response to signed order flow (order size), and by the fixed cost of trading using intra-day continuous data on transactions and quotes. Amihud (2002) measures a stock's illiquidity as the ratio of its absolute return to dollar volume. Pastor and Stambaugh (2003) estimate a liquidity risk measure based on the idea that price changes accompanying large volume tend to be reversed when market-wide liquidity is low.

Among these liquidity proxies, Amihud's illiquidity measure (Amihud, 2002) is widely used in empirical studies because of its superior advantage of simple calculation. In addition, this proxy only needs return and volume data so that we can estimate liquidity for a relatively long time-span. Amihud's measure is also consistent with Kyle's (1985) concept of illiquidity, the response of price to the order flow, and Silber's (1975) measure of thinness, which is defined as the ratio of the absolute price change to the absolute excess demand for trading. The liquidity measure for stock i is defined as:

$$\gamma_{i,m} = -\frac{1}{D_{i,m}} \sum_{t=1}^{D_{i,m}} \frac{\left| r_{i,m,d} \right|}{v_{i,m,d}} \tag{9.2}$$

where $D_{i,m}$ is the number of days for which data are available for stock i in month m, $R_{i,m,d}$ is the return on stock i on day d of month m, and $v_{i,m,d}$ is the dollar trading volume for stock i on day d of month m.

However, this measure has two disadvantages; the first of which is that Amihud (2002) takes dollar trading volume as the denominator, which may result in a high correlation between liquidity and size since large stocks are usually more frequently traded than small stocks. We do not expect the same dollar amount of trading for a firm whose market capitalization is 10 million dollars as for a firm whose market capitalization is 10 billion dollars. In addition, as share prices increase over time, liquidity appears to increase when it is measured by Amihud's method even if there are no changes in liquidity. In order to construct a liquidity measure that is robust to size, we scale the denominator by the market capitalization of the stock; in other words, dollar trading volume is replaced by turnover rate in the denominator. Furthermore, as argued by Lo and Wang (2000), turnover is a canonical measure of trading activity. Therefore, while replacing the dollar trading volume by turnover does not alter the principle of this price reversal nature, it enables us to construct a relative liquidity proxy that is free from size effect.

The second reason is that the liquidity measure may have severe outliers when trading activity is extremely low (i.e., trading volume could be very close to zero); therefore, we use the natural log of these values to minimize these outliers. The modified relative liquidity measure ($\psi_{i,m}$) for stock i is defined as:

$$\psi_{i,m} = -\frac{1}{D_{i,m}} \sum_{t=1}^{D_{i,m}} \ln \frac{|r_{i,m,d}|}{Turnover_{i,m,d}} \tag{9.3}$$

where $D_{i,m}$ is the number of days for which data are available for stock i in month m, $R_{i,m,d}$ is the return on stock i on day d of month m, and $Turnover_{i,m,d}$ is the turnover rate for stock i on day d of month m.

The liquidity measures of equations (9.2) and (9.3) are calculated for all stocks in every month, from which we obtain the monthly liquidity measures for each stock. We also create two market-wide liquidity factors using a similar method to that in Fama and French (1993). At the end of each year, two portfolios are formed on liquidity using the median liquidity as the breakpoint. The return difference of the liquid and illiquid portfolios in the following 12 months is the liquidity factor (LIQ_t). The portfolios are rebalanced at the end of each year. We calculate each stock's liquidity and market-wide liquidity factor for Amihud's (2002) illiquidity measure ($\gamma_{i,m}$) and size-adjusted liquidity measure ($\psi_{i,m}$). With this mimicking liquidity factor, we can explore the role of liquidity in explaining the cross-sectional return differences.

Size and value/growth factors

UK *SMB* and *HML* factor returns are calculated in a similar way to that in Fama and French (1993) except for the breakpoints. Fama and French (1993) use the fiftieth percentile (for size) and thirtieth and seventieth percentiles (for book-to-market) NYSE-based breakpoints. Following Dimson *et al.* (2003b), however, we use the seventieth percentile of ranked size and fortieth and sixtieth percentiles of book-to-market. In the UK, large-capitalization stocks are concentrated in the low book-to-market segment and small-capitalization stocks, in contrast, are concentrated in the high book-to-market class

(similar results are found in this study, and are displayed in Section 9.3). By choosing less extreme book-to-market breakpoints and a wider range for the small-capitalization group, it ensures acceptable levels of diversification in these corner portfolios throughout the sample period. In addition, the 70% breakpoint for the size results in a distribution of aggregate market value across portfolios that is relatively similar to the distribution in Fama and French (1993), where most NASDAQ stocks, most of which are smaller than the NYSE-based 50% breakpoint, are sorted into the small-capitalization group.

9.3 Empirical results

9.3.1 Data

In this study we use the sample period starting from January 1987 to December 2004, because the trading volume data which we use for calculating liquidity are not available pre-1987. All the data on stock returns, market capitalizations, book-to-market ratios and trading volumes are from DataStream. In calculating liquidity measures, the daily frequency returns and dollar trading volume data are needed. The book-to-market ratio is the end of the calendar year value, and any negative book-to-market stocks are deleted from the sample; in addition, all the delisted equities are also included in the sample so that survivorship bias is controlled for in this study. Initially the number of stocks in this study is 945, in 1987; this gradually increases to 2306 in 2004. While calculating the liquidity measures, due to the lack of availability of data regarding trading volume, the sample size reduces to less than 200 stocks from 1987 to 1990, 576 stocks in 1991 and 1459 stocks in 2004 (the number of stocks used to calculate liquidity measures and factors are displayed in the final row of Table 9.1). For the time-series regression analysis in our study, we choose a sample period from January 1991 to December 2004 in order to minimize any bias that may arise from the small number of stocks in our early sample period.

9.3.2 Market liquidity

In a manner that is consistent with Amihud (2002) and Pastor and Stambaugh (2003), market illiquidity is defined as the average of individual stock's illiquidity. The first two rows of Table 9.1 display the value- and equally-weighted Amihud's illiquidity measure. The value-weighted Amihud's illiquidity measure suggests that the most illiquid year is 2001, which corresponds with 11 September; while 1998 is the second lowest liquidity period, corresponding with the Russian Crisis.[3] The equally-weighted Amihud's illiquidity also shows similar patterns; however, there the results are relatively more skewed towards small stocks in the market, which thus is more volatile, suggesting that the level of illiquidity of small stocks changes more than that of large stocks. Figure 9.1 clearly shows that the illiquidity measures are associated with market crashes – for instance, the 1997 Asian financial crisis, the 1998 Russian default, 11 September 2001 etc.

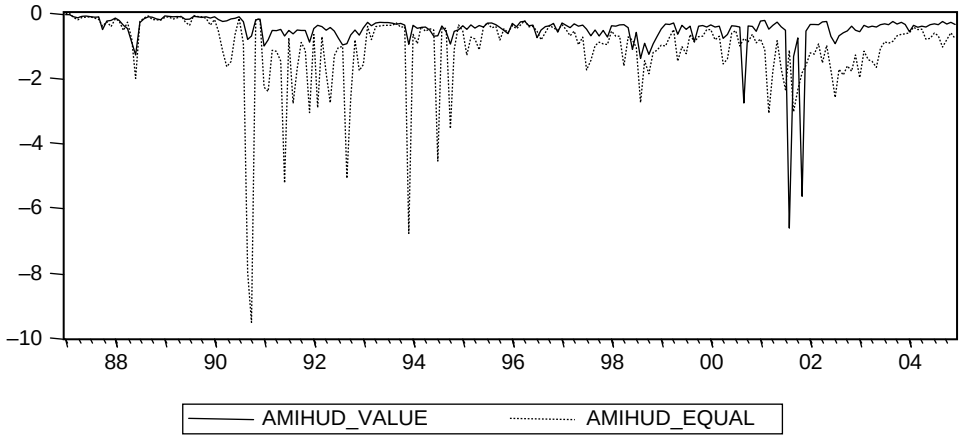
The third row of Table 9.1 shows the relative liquidity measure (adjusted by stocks' market values), described in Section 9.2. Figure 9.2 plots our relative liquidity measure.

[3]In contrast, the most liquid year is 1989; however, the high liquidity in the early sample period is likely a result of the sample selection bias in the early sample period.

Table 9.1 Liquidity properties of the UK stock market over time (annual average)

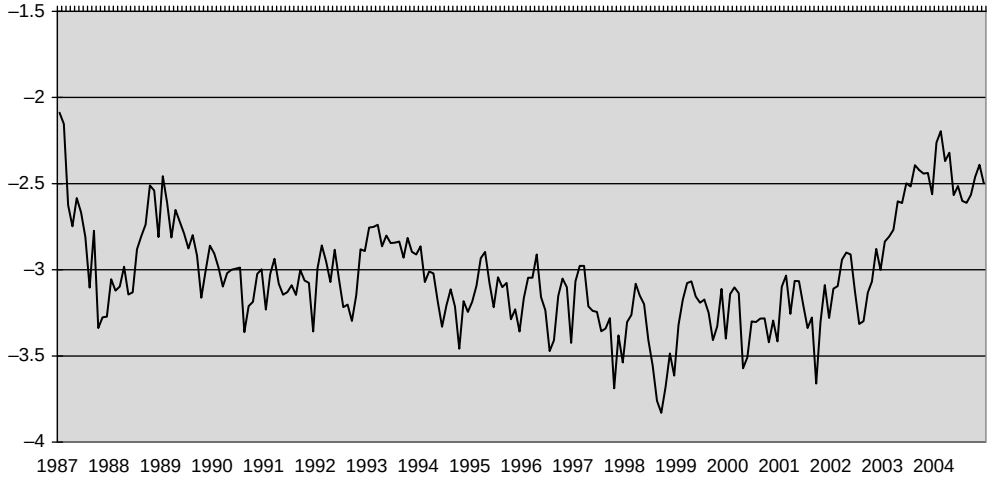
	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
$\gamma_{i,m}$ (value weighted)	−0.16	−0.38	−0.13	−0.29	−0.65	−0.62	−0.39	−0.56	−0.43	−0.42	−0.50	−0.75	−0.46	−0.67	−1.46	−0.51	−0.38	−0.38
$\gamma_{i,m}$ (equally weighted)	−0.18	−0.43	−0.18	−2.06	−2.01	−1.88	−1.06	−1.18	−0.67	−0.47	−0.91	−1.22	−0.88	−0.87	−1.81	−1.56	−1.13	−0.68
$\psi_{i,m}$	−2.79	−2.90	−2.81	−3.06	−3.11	−3.04	−2.83	−3.16	−3.12	−3.18	−3.28	−3.44	−3.22	−3.31	−3.22	−3.06	−2.57	−2.45
No. of stocks	96	179	183	184	576	577	712	826	832	798	974	1003	1060	1118	1325	1367	1382	1459

The market liquidity is the average of individual stock's illiquidity. The first two rows are aggregate market liquidity based on Amihud's (2002) measure. The third row is the aggregate market liquidity based on our new relative liquidity measure.



Note: The market liquidity is the average of individual stock's illiquidity, where the illiquidity measure is based on Amihud's (2002) measure.

Figure 9.1 Monthly average Amihud's illiquidity level in the UK.



Note: The market liquidity is the average of individual stock's illiquidity, where the liquidity measure is based on our new relative measure.

Figure 9.2 Monthly average relative illiquidity level in the UK 1987 to 2004.

While the monthly correlation between this relative measure and the value (equally)-weighted Amihud's measure is 0.29 (0.25), it is interesting to note that there is a significant difference between Amihud's absolute illiquidity measure and the new relative measure. The new relative measure is much smoother and less volatile than Amihud's measure. Severe outliers in Figure 9.1 are now apparently reduced according to this new liquidity measure. With this measure, we can clearly identify that the most illiquid year is 1998, when market liquidity is widely perceived to have dried up because of the LTCM collapse

and Russian default.[4] The next illiquid periods are the Asian financial crisis of 1997, and 11 September 2001. By contrast, the most liquid period is during the recent bull market. Amihud (2002) shows liquidity displays persistence, and indeed the new measure has first- and second-order autocorrelations of 0.83 and 0.72, which are both significant at 1% level. From Figure 9.2 it is evident that liquidity remains at a relative low level in the late 1990s, when the market is down, and recovers gradually with the recent bull market.

9.3.3 Different portfolio strategies in the UK

The existence of return regularities is well documented in the financial markets. Previous work shows that average stock returns are related to firm characteristics like size and book-to-market equity. In this section, we examine the cross-sectional stock returns related to size, book-to-market equity and liquidity in the UK. We also compare the results to the US, as in Fama and French (1993), and the previous work on the UK market in Dimson *et al.* (2003b).

Size-sorted portfolios

As described in the previous section, the ten decile portfolios are formed based on market equity. Table 9.2a shows the statistical properties of the ten decile portfolios. Contrary to the findings of Banz (1981) and Fama and French (1993), where they find evidence that small firms outperform big firms, it is interesting to see that big firms perform better than small firms in the UK equity market (there is a 5.7% annual difference in the portfolio returns between the largest and smallest decile portfolios, i.e. S–B_d).

It is evident that the largest 10% of stocks represent, on average, 81% of the total market capitalization (the largest 20% of stocks account for over 90% of the total market capitalization). During the same period in the US market, the largest 20% of stocks account for about 80% of the total market capitalization.[5] This suggests a more skewed distribution of large stocks in the UK stock market.

Table 9.2a Properties for ten decile-size portfolios (percentage) January 1988–December 2004

	Small	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Big**	***SMB***
Average MV/total MV (%)	0.06	0.16	0.30	0.49	0.81	1.32	2.24	4.05	9.29	81.29	
Average annual return	1.31	0.89	0.50	0.81	2.60	2.50	2.34	2.79	4.00	7.05	**−4.08**
Average monthly return	0.11	0.07	0.04	0.07	0.21	0.21	0.19	0.23	0.33	0.57	**−0.35**
Monthly return SD	3.32	3.52	3.51	3.81	4.07	3.89	4.00	3.91	4.04	3.90	**2.72**

At the end of June in year *t*, ten size-decile portfolios are formed on stocks' ranked market value and held for the next twelve months. Every portfolio represents 10 percentile of the ranked ME. Portfolios are rebalanced every year. *SMB* is the mimicking factor for size. The breakpoint is the seventieth percentile of the ranked market equity.

[4]Pastor and Stambaugh (2003), who use the US data and their proposed price-reversal liquidity measure, identify that the US stock market experience the third largest liquidity drop in 1998. Within the same time-span, however, our study shows consistent results with theirs.

[5]The US data over this period are from Professor Kenneth French's website.

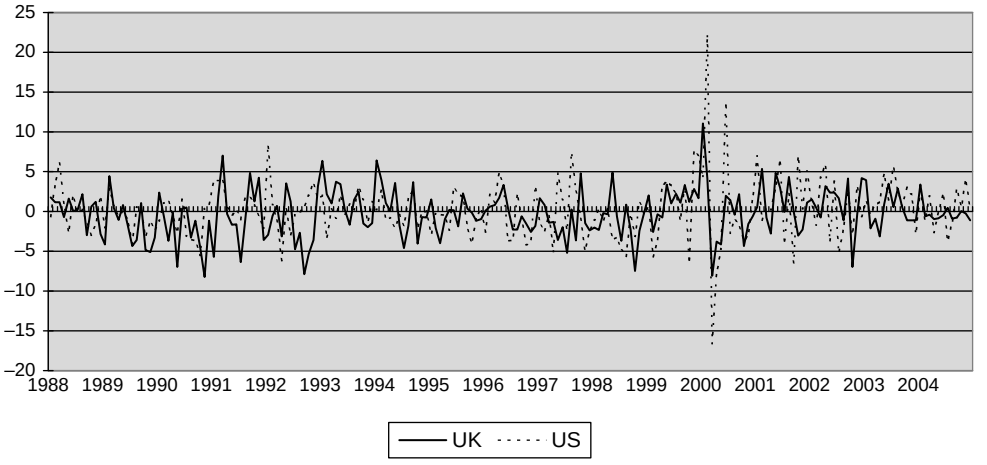
Table 9.2b Statistical properties for monthly *SMB* in the UK and US

	SMB_US	SMB_UK
Mean	0.20%	−0.35%
Median	0.20%	−0.40%
SD	3.70%	2.70%
t-test	0.66	−0.25

The data for the US is from Kenneth French's website.

We next calculate the mimicking size factor (*SMB*) for the UK stock market. The statistical properties of the *SMB* are reported in the last column of Table 9.2a. Over this entire 17-year period, the *SMB* has a negative average monthly return of 0.35% (which is equal to an annual average return of −4.08%) with standard deviation of 2.7%. Our results for the UK market are consistent with the findings of Dimson *et al.* (2003b), where, although the data in their research only go up to 2001, the correlation of their monthly *SMB* and ours is nearly 92%.

By contrast, the *SMB* is positive with an average monthly return of 0.2% in the US over this period. Nevertheless, the trends for the *SMB* in the UK and US are very similar, as shown by Figure 9.3. Indeed, the annual (monthly) *SMB* between the UK and US has a correlation of 0.70 (0.33). Table 9.2b compares the statistical properties of monthly *SMB* in the UK and US. The *t*-tests with regard to a zero mean for the *SMB* suggest that both the UK and US *SMB* is insignificantly different from zero. Ang and Chen (2005)



The UK *SMB* is calculated as follows: at the end of June in year *t*, two portfolios are formed on stocks' ranked market value and hold for the next 12 months. The breakpoint is the 70th percentile of the ranked market equity. The return difference of these two portfolios is the *SMB*. Portfolios are rebalanced every year. The US *SMB* is downloaded from Kenneth French's website.

Figure 9.3 Monthly *SMB* in the UK and US from 1988 to 2004 (percentage).

and Dimson and Marsh (1999) also document the disappearance of the size effect in the US and UK respectively.

Book-to-market-sorted portfolios

Table 9.3a reports the summary statistics for the ten decile book-to-market portfolios. The annual return difference between the high and low book-to-market portfolios (H–L_D) is over 10%.[6] Consistent with Fama and French (1998), and Dimson *et al.* (2003b), there is strong evidence of the existence of the book-to-market premium. The results also show that small stocks are usually distributed in the high book-to-market category. Half of the highest book-to-market stocks only account for 20% of the total market capitalization. These results are consistent with the finding of Fama and French (1993) and Dimson *et al.* (2003).

The mimicking value/growth factor (*HML*) is reported in the last column of Table 9.3a. With a standard deviation of 2.5%, the monthly *HML* has a return of 0.32% (which equals an annual rate of 3.9%). In the US, these numbers are larger (the annual *HML* is 4.9%, with a standard deviation of 3.4%). There is again a similar trend in the *HML* during the same period in the US and UK, which can be seen from Figure 9.4. The correlation of the annual (monthly) *HML* between UK and US is 0.24 (0.15). Table 9.3b reports the statistical comparison of the monthly *HML* in the UK and US. The zero-mean tests suggest that both of them are significantly different from zero.

Table 9.3a Properties for ten decile-book-to-market portfolios (percentage) January 1988–December 2004

	Low	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**High**	**HML**
Average MV/total MV (%)	17.28	14.56	17.30	16.53	9.96	7.74	5.53	4.07	3.13	3.90	
Average annual return	2.08	1.34	6.08	7.49	4.22	6.93	6.12	8.02	9.62	12.27	**3.87**
Average monthly return	0.17	0.11	0.49	0.60	0.34	0.56	0.50	0.64	0.77	0.96	**0.32**
Monthly return SD	3.83	4.48	3.93	4.21	4.36	4.68	5.46	5.03	4.83	4.94	**2.50**

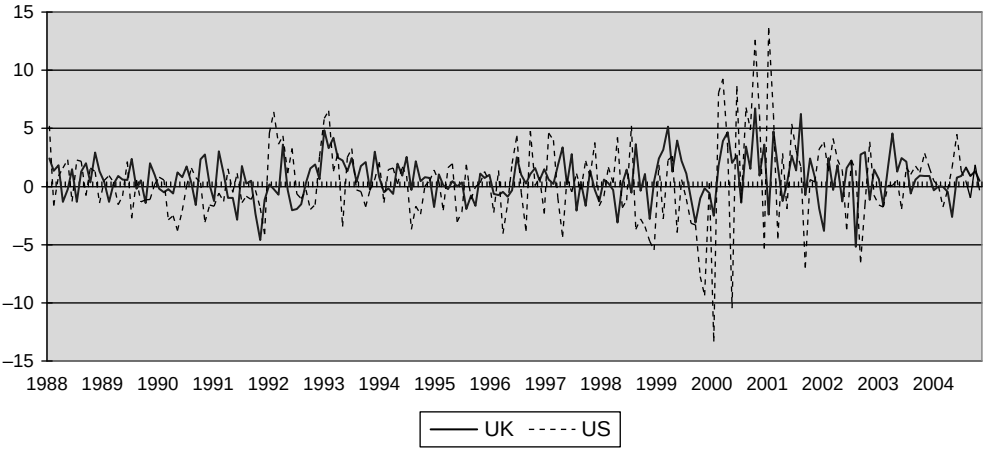
At the end of December in year *t*, ten value-decile portfolios are formed on stocks' ranked book-to-market value and held for the next twelve months. Every portfolio represents 10 percentile of the ranked book-to-market stocks. Portfolios are rebalanced every year. HML is the mimicking factor for value. The breakpoints are the fortieth and sixtieth percentiles of the ranked book-to-market equity.

[6]Fama and French (1998) find that there is a value premium of 4.62%. This is because they use a very small sample for the UK market, where on average only 185 stocks are examined.

Table 9.3b Statistical properties for monthly *HML* in the UK and US

	HML_US	HML_UK
Mean	0.30%	0.32%
Median	0.30%	0.20%
SD	3.40%	2.50%
t-test	4.42	4.52

The data for the US is from Kenneth French's website.



The UK *HML* is calculated as follows. At the end of December in year *t*, two portfolios are formed on stocks' ranked book-to-market value and hold for the next 12 months. The breakpoints are the 40th and 60th percentile of the ranked book-to-market equity. The return difference of these two portfolios is the *HML*. Portfolios are rebalanced every year. The US *HML* is downloaded from Kenneth French's website.

Figure 9.4 Monthly *HML* in the UK and US 1988 to 2004 (percentage).

Liquidity-sorted portfolios

Based on the illiquidity measure of Amihud (2002), ten decile liquidity portfolios are formed and summarized in Table 9.4a. The risk-return relationship suggests that illiquid stocks should earn higher expected returns than liquid stocks, because investors should be compensated for bearing the illiquidity risk. However, Table 9.4a indicates that this is not the case in the UK equity market. On average, the highly illiquid 30% of the stocks display a negative annual return from 1988 to 2004. The most illiquid-decile equity group in the UK even experienced a −13.7% annual loss. In contrast, the most liquid-decile stocks show an annual return of nearly 9%, which results in an annual return spread of over 22% between the liquid and illiquid stocks (ILLIQ–LIQ_d).

The three most illiquid portfolios include many small stocks, where the total market value of these portfolios is only 2.59%. As liquidity increases, so does the size of the

Table 9.4a Properties for ten decile-liquidity portfolios sorted by $\gamma_{i,m}$ 1988 to 2004 (percentage)

	Illiquid	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Liquid**	***LIQ***
Average MV/total MV (%)	0.74	0.78	1.07	1.36	1.73	2.34	3.17	5.15	11.19	72.47	
Average annual return	−13.74	−2.67	−3.44	1.76	1.75	−0.03	1.88	7.03	3.64	8.74	**−7.85**
Average monthly return	−1.23	−0.23	−0.29	0.15	0.14	0.00	0.16	0.57	0.30	0.70	**−0.68**
Monthly return SD	6.86	5.65	5.45	5.23	5.36	5.20	5.19	5.03	5.29	4.45	**2.92**

At the end of December in year t, ten liquidity-decile portfolios are formed on stocks' ranked with Amihud's illiquidity measure and held for the next twelve months. Every portfolio represents 10 percentile of the ranked stocks. Portfolios are rebalanced every year. *LIQ* is the mimicking liquidity factor, which is the average return difference between the portfolios of the most illiquid 50% and the most liquid 50% of stocks.

Table 9.4b Properties for ten decile-liquidity portfolios sorted by $\psi_{i,m}$ 1988 to 2004 (percentage)

	Illiquid	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Liquid**	***LIQ***
Average MV/total MV (%)	9.01	6.57	9.26	10.79	13.48	15.80	13.08	10.87	8.59	2.56	
Average annual return	−6.55	−0.98	3.58	4.11	6.01	9.10	7.27	6.80	11.96	11.37	**−7.53**
Average monthly return	−0.56	−0.08	0.29	0.34	0.49	0.73	0.59	0.55	0.94	0.90	**−0.65**
Monthly return SD	6.09	4.98	4.69	5.16	4.85	4.63	4.37	5.07	5.18	7.03	**2.02**

At the end of December in year t, ten liquidity-decile portfolios are formed on our ranked relative liquidity measure and held for the next twelve months. Every portfolio represents 10 percentile of the ranked stocks. Portfolios are rebalanced every year. *LIQ* is the mimicking liquidity factor, which is the average return difference between the portfolios of the most illiquid 50% and the most liquid 50% of stocks.

firms, where the most liquid 10% of the stocks stand for over 72% of the total market capitalization. As expected, Amihud's illiquidity measure is affected by the size of firms, where large firm's stocks tend to be more liquid than those for small firms. From Table 9.2a we see that big stocks have an average annual return of 7.05%, which is very similar to the return for the liquid stocks. Contrary to this, because the illiquid stocks (usually small) show negative returns, we can infer that small stocks with low liquidities perform much worse.

While Amihud's illiquidity measure is apparently highly correlated with size,[7] as expected, our relative liquidity measure should not be. Summary statistics of the ten portfolios, made upon the relative liquidity measure, are presented in Table 9.4b. The first row reports the percentage of each liquidity-decile's market capitalization to the total market capitalization. Although this time the most liquid stock group shows a smaller weight than other groups, the remaining nine deciles are much more evenly distributed

[7]In Table 9.5a, we can see the correlation between the liquidity mimicking factor based on Amihud's measure and the *SMB* is 75%.

in terms of size. Therefore, small stocks are illiquid in absolute measure, but they could be as liquid as (or more liquid than) larger stocks according to our relative measure.

However, the return spread (ILLIQ–LIQ_d) based on this relative liquidity measure is still large (almost 18% annually), although it is smaller than that based on Amihud's measure. The standard deviation for the most liquid portfolio is larger than that of Amihud's measure. This could be a result of the fact that the liquid stocks based on Amihud's measure are usually large, and their returns are less volatile.

The monthly correlation matrix between the *RM, SMB, HML* and the two liquidity mimicking factors is displayed in Table 9.5a. The mimicking liquidity factor based on our relative liquidity measure barely shows any relationship with that based on Amihud's measure, and also very low correlation with *SMB* and *RM*. Amihud's measure is, however, highly correlated with size. There is also almost no relationship between the *SMB* and the *HML* in the UK, a result similar to the US (Fama and French, 1993). Table 9.5b describes the statistical properties of these factors. Because liquid and big stocks display excess returns over illiquid and small stocks in the UK, this makes the mimicking liquidity factor (*LIQ*) and size factor (*SMB*) display negative values. The last row of Table 9.5b reports the Sharpe ratios of various factors. *LIQ* and *LIQ_AMIHUD* produce the largest absolute Sharpe ratios, which implies that investors can be significantly rewarded for perusing the liquidity strategies – specifically, buying liquid and selling illiquid stocks.

Table 9.5a Correlation matrix (204 monthly observations)

	LIQ	*LIQ_AMIHUD*	*RM*	*HML*	*SMB*
LIQ	1				
LIQ_AMIHUD	0.01	1			
RM	−0.08	−0.12	1		
HML	−0.24	0.03	0.20	1	
SMB	−0.08	0.75	−0.33	0.02	1

Table 9.5b Statistical properties for all factors in the UK

	LIQ	*LIQ_AMIHUD*	*RM*	*HML*	*SMB*
Monthly mean	−0.65%	−0.68%	0.80%	0.32%	−0.35%
Monthly SD	2.02%	2.92%	4.20%	2.50%	2.72%
Historical Sharpe ratio	−0.65	−0.46	0.03	0.01	−0.31

RM is the market return on the FTSE-all share index. *SMB* is the return difference between the seventieth percentile of the ranked *ME* at the end of June each year *t*, and *HML* is the return spread between the fortieth and sixtieth percentiles of (*BE/ME*) at the end of December each year *t*. At the end of each year, Amihud's illiquidity measure and our relative liquidity measure are calculated for all stocks. The two portfolios are created based on these ranked measures. The breakpoint is the median of each liquidity measure. The return difference of these two portfolios in the next twelve months is the mimicking liquidity factor (*LIQ* and *LIQ_AMIHUD*).

Long big and short small also tends to be a good investment strategy in the UK stock market, as it has the third largest Sharpe ratio.

9.3.4 Liquidity effects in explaining cross-sectional returns

From the previous section, we can see that average returns are closely related to the stock characteristics, such as size, book-to-market and liquidity. In this section, we examine the cross-sectional effect of liquidity on stock returns.

We first test the CAPM on different portfolios strategies related to size, book-to-market and liquidity. The statistical properties of different hedge portfolios are reported in Table 9.6a. It is evident that the strategy of small minus big displays negative average returns regardless of breakpoints. However, the *t*-tests suggest that these negative values are not statistically significant. The book-to-market strategy, by contrast, shows a significantly positive average return whatever breakpoint is employed. Liquid stocks have significant excess returns over illiquid stocks, as highlighted by the right-hand panel of the table.

Table 9.6b reports the results of the CAPM on different hedge portfolios related to size, book-to-market and liquidity. It is evident that the CAPM itself is valid for the three size strategies, as there is no significant alpha in these three regressions. By contrast, there are significant value and liquidity anomalies in the UK market. The alphas are all significantly different from zero, which suggests the failure of the CAPM in explaining this book-to-market and liquidity return regularities.

Table 9.7 describes the effect of liquidity on the size and book-to-market strategies. Although the CAPM is efficient in explaining the return regularities associated with size, as shown in Table 9.6b, it is still of interest to note that liquidity betas are all significant at the 95% confidence level in the left-hand panel of Table 9.7. This evidence implies that the liquidity risk partly explains the excess return of big over small stocks. The right panel of Table 9.7 shows that all the factor loadings on liquidity are statistically significant at the 99% confidence level, which suggests that liquidity plays a significant role in describing the observed value anomaly. In this liquidity-augmented CAPM, all three book-to-market strategies show insignificant intercepts. Compared with the central panel of Table 9.6b, the magnitude of intercepts are dramatically reduced in this liquidity-augmented model, which illustrates the success of the liquidity factor in explaining value anomalies.

Fama and French (1995) argue that the *HML* proxies for relative financial distress risk. Saretto (2004) provides empirical evidence that *HML* can be interpreted as a distress factor. Chen and Zhang (1998) also demonstrate that the high returns from value stocks compensate for the high risks induced by the characteristics such as financial distress, earnings uncertainty or financial leverage.

A seminar work by Agarwal and Taffler (2005) uses a z-score as a proxy for distress risk, and shows that momentum is largely subsumed by their distress risk factor. We investigate whether the role of liquidity in explaining value premium is not subsumed by their distress factor. The distress risk factor is calculated using the same method described by Agarwal and Taffler (2005).[8] The correlation between our mimicking liquidity factor (LIQ_t) and the distress factor is -0.09. The financial distress factor cannot explain the

[8]We are grateful to Vineet Agarwal for the generous provision of their UK financial distress factor.

Table 9.6a Monthly statistical properties of different hedge portfolios

	S–B			H–L			aILLIQ–LIQ		
	S–B_D	S–B_Q	S–B_P	H–L_D	H–L_Q	H–L_P	ILLIQ–LIQ_D	ILLIQ–LIQ_Q	ILLIQ–LIQ_P
Mean	−0.271%	−0.257%	−0.251%	0.819%	0.756%	0.544%	−0.914%	−1.243%	−1.462%
SD	3.809%	3.397%	3.108%	4.366%	3.266%	2.670%	3.223%	4.082%	6.408%
t-test	−0.922	−0.980	−1.046	2.431	3.001	2.641	−3.259	−4.349	−4.048

S–B (H–L) is the hedge portfolio for long the smallest (highest book-to-market) and short the biggest (lowest book-to-market). ILLIQ–LIQ is the hedge portfolio for long the most illiquid and short most liquid. _D, _Q and _P stand for decile, quintile and 30%/40%/30% breakpoints.

Table 9.6b CAPM of SMB, HML and ILLIQ–LIQ from 1991 to 2004

	S–B			H–L			aILLIQ–LIQ		
	S–B_D	**S–B_Q**	**S–B_P**	**H–L_D**	**H–L_Q**	**H–L_P**	**ILLIQ–LIQ_D**	**ILLIQ–LIQ_Q**	**ILLIQ–LIQ_P**
C	−0.001	−0.001	−0.001	0.008	0.008	0.005	−0.010	−0.010	−0.007
	−0.346	*−0.443*	*−0.508*	*2.028*	*2.402*	*1.956*	*−2.183*	*−3.405*	*−3.123*
RM-Tbill	−0.546	−0.439	−0.395	0.047	0.004	0.068	−0.461	−0.330	−0.276
	−7.988	*−6.807*	*−6.659*	*0.382*	*0.036*	*0.862*	*−3.446*	*−3.567*	*−3.981*
R-squared	0.348	0.282	0.273	0.002	0.000	0.011	0.114	0.126	0.135

RM is the market return on the FTSE-all share index. Tbill is the monthly rate for the UK one-month Treasury bill. S–B (H–L) is the hedge portfolio for long the smallest (highest book-to-market) and short the biggest (lowest book-to-market). ILLIQ–LIQ is the hedge portfolio for long the most illiquid and short the most liquid. _D, _Q and _P stand for the decile, quintile and 30%/40%/30% breakpoints. The numbers in italic are *t*-statistics. Portfolios are rebalanced every year. The estimations are justified in the presence of both heteroscedasticity and autocorrelation of unknown forms according to Newey and West (1987).

Table 9.7 Liquidity effects over size (*SMB*) and book-to-market strategy (*HML*)

	S–B			H–L		
	S–B_D	**S–B_Q**	**S–B_P**	**H–L_D**	**H–L_Q**	**H–L_P**
C	−0.003	−0.003	−0.003	0.005	0.005	0.003
	−1.146	*−1.151*	*−1.307*	*1.503*	*1.609*	*1.083*
RM-Tbill	−0.586	−0.474	−0.432	−0.006	−0.053	0.014
	−8.315	*−6.957*	*−7.170*	*−0.051*	*−0.498*	*0.184*
LIQ	−0.358	−0.321	−0.343	−0.489	−0.511	−0.488
	−2.502	*−2.061*	*−2.664*	*−2.859*	*−4.073*	*−4.994*
Wald test of LIQ (F-stats)	6.261	4.246	7.095	8.176	16.591	24.937
R-squared	0.381	0.315	0.318	0.049	0.091	0.136

RM is the market return on the FTSE-all share index. Tbill is the monthly rate for the UK one-month Treasury bill. S–B (H–L) is the hedge portfolio for long the smallest (highest book-to-market) and short the biggest (lowest book-to-market). _D, _Q and _P stand for the decile, quintile and 30%/40%/30% breakpoints. LIQ is the liquidity mimicking factor based on our size-adjusted liquidity measure. The numbers in italic are *t*-statistics. Portfolios are rebalanced every year. The estimations are justified in the presence of both heteroscedasticity and autocorrelation of unknown forms according to Newey and West (1987).

value premium in the UK market. Table 9.8a describes a distress-factor augmented CAPM, where it is clear that the CAPM intercepts remain significant and factor loadings on distress factor are all insignificantly different from zero. In Table 9.8b, when the distress factor is added into the liquidity-augmented CAPM, the relation between liquidity and value premium continues to be significant.

The possible explanation for the close relationship between liquidity and value premium can be found in Campbell *et al.* (1993), where they present a model in which trading volume proxies for the aggregate demand of liquidity trading. In addition, the empirical work of Lee and Swaminathan (2000) demonstrates that low (high) volume stocks display many characteristics commonly associated with value (growth) stocks. Therefore, the return spread between value and growth could contain the difference in the liquidity risk inherited by them. Thus liquidity could help to explain this value premium.

Table 9.8a Financial distress factor over book-to-market strategy

H–L	**H–L_D**	**H–L_Q**	**H–L_P**
C	0.009	0.008	0.006
	1.954	*2.349*	*2.026*
RM-Tbill	0.074	−0.012	0.039
	0.566	*−0.108*	*0.511*
Distress_factor	−0.096	0.137	0.184
	−0.468	*0.871*	*1.451*
R-squared	0.005	0.007	0.028

Table 9.8b Robustness of liquidity effects over book-to-market strategy

H–L	H–L_D	H–L_Q	H–L_P
C	0.006	0.005	0.003
	1.369	*1.563*	*1.193*
RM-Tbill	0.027	−0.056	−0.002
	0.206	*−0.524*	*−0.029*
LIQ	−0.543	−0.520	−0.480
	−3.150	*−3.970*	*−4.791*
Distress_factor	−0.173	0.064	0.117
	−0.914	*0.477*	*1.109*
Wald test of LIQ (F-stats)	9.922	15.761	22.952
R-squared	0.059	0.094	0.143

RM is the market return on the FTSE-all share index. Tbill is the monthly rate for the UK one month Treasury bill. H–L is the hedge portfolio for long the highest book-to-market and short the lowest book-to-market. _D, _Q and _P stand for the decile, quintile and 30%/40%/30% breakpoints. LIQ is the liquidity mimicking factor based on our size-adjusted liquidity measure. The distress factor is a mimicking factor for distress risk, obtained from Agarwal and Taffler (2005). The numbers in italic are *t*-statistics. Portfolios are rebalanced every year. The estimations are justified in the presence of both heteroscedasticity and autocorrelation of unknown forms according to Newey and West (1987).

9.3.5 Robustness of liquidity effects

In this section, we first demonstrate the result that liquidity can be employed to explain the value premium in the UK is robust to a variety of macroeconomic variables. Secondly, we argue that our liquidity factor is robust in the sense that the cross-sectional return difference related to liquidity is unexplainable by other well-known factors, such as, *SMB, HML* and momentum.

Zhang (2005) proposes that the value premium is linked to macroeconomic conditions. He explains that value firms are burdened with more unproductive capital when the economy is bad, and find it more difficult to reduce their capital stock than growth firms do. The dividends and returns of value stocks will hence co-vary more with economic downturns. We demonstrate that the ability of liquidity to explain the value premium is robust in the presence of a battery of macroeconomic variables,[9] such as industrial production, CPI, money supply, term spread (i.e., yield difference between 10-year government bond and 1-month Tbill), and corporate spread (i.e., the yield difference between BBB and AAA bonds). Such results are displayed in Table 9.9.

Table 9.10 suggests that the liquidity premium remains pronounced in different models, such as the CAPM, the Fama and French Model, the Fama and French model augmented

[9]The data from these macroeconomic variables is from OECD; refer to the footnote of Table 9.9 for details.

Table 9.9 Robustness of liquidity effects over macroeconomic variables

HML	H–L_D	H–L_Q	H–L_P
C	0.013	0.015	0.005
	1.664	*2.753*	*1.051*
RM-Tbill	−0.027	−0.140	−0.088
	−0.131	*−0.863*	*−0.784*
LIQ	−0.343	−0.332	−0.389
	−1.878	*−2.440*	*−3.580*
Industrial production	−0.191	0.260	0.116
	−0.382	*0.743*	*0.412*
CPI	1.225	0.582	0.400
	0.782	*0.495*	*0.444*
Term spread	7.261	3.514	2.147
	1.113	*0.645*	*0.440*
Corporate spread	−0.538	−0.186	0.271
	−0.612	*−0.296*	*0.590*
Money supply (M2)	−1.084	−1.206	−0.358
	−1.344	*−1.964*	*−0.613*
Wald test of LIQ(F-stats)	3.527	5.953	12.816
R-squared	0.051	0.106	0.118

RM is the market return on the FTSE-all share index. Tbill is the monthly rate for the UK one month Treasury bill. H–L is the hedge portfolio for long the highest book-to-market and short the lowest book-to-market. _D, _Q and _P stand for the decile, quintile and 30%/40%/30% breakpoints. LIQ is the liquidity mimicking factor based on our size-adjusted liquidity measure. The macroeconomic variables are all downloaded from Datastream. Among these macroeconomic variables, the industrial production is the increase rate of the seasonally adjusted UK industrial production volume index. CPI is the changes of the UK consumer price index. Term spread is the yield difference between the ten-year government bond and one-month Tbill. Corporate spread is the return difference between the Merrill Lynch UK BBB and AAA bond index. Money supply is the increased rate of the broad money supply. The numbers in italic are *t*-statistics. Portfolios are rebalanced every year. The estimations are justified in the presence of both heteroscedasticity and autocorrelation of unknown forms according to Newey and West (1987).

with a distress factor, and the Fama and French model augmented with a momentum factor (Winners minus Losers).[10]

In short, the finding that the value premium is related to liquidity is robust to a number of macroeconomic variables. The premium of liquid over illiquid stocks is robust even after adjusted by *SMB, HML* distress and momentum factors.

[10]We report the case of decile breakpoints simply for reasons of presentation. However, the results are similar regardless of different strategies such as ILLIQ–LIQ_q or ILLIQ–LIQ_p.

Table 9.10 Different factor models for illiquid minus liquid portfolios

Dependent variables: ILLIQ–LIQ_D					
C	−0.010	−0.012	−0.014	−0.0147	−0.016
	−2.183	*−2.543*	*−2.622*	*−2.702*	*−2.846*
RM-Tbill	−0.461	−0.614	−0.552	−0.5595	−0.504
	−3.446	*−4.845*	*−4.506*	*−3.844*	*−3.730*
SMB		−0.478	−0.361	−0.4354	−0.338
		−3.776	*−2.499*	*−2.933*	*−2.109*
HML		0.401	0.444	0.4119	0.452
		2.216	*2.443*	*2.2824*	*2.497*
Distress_factor			−0.227		−0.145
			−0.973		*−0.583*
WML				0.1566	0.179
				1.091	*1.134*
R-squared	0.114	0.190	0.198	0.197	0.207

RM is the market return on the FTSE-all share index. Tbill is the monthly rate for the UK one-month Treasury bill. ILLIQ–LIQ_D is the return difference between the most illiquid and liquid decile portfolios. *SMB* and *HML* are mimicking factors for size and value. The distress factor is a mimicking factor for distress risk, obtained from Agarwal and Taffler (2005). WML is a six by six momentum factor as in Jegadeesh and Titman (1993). The numbers in italic are *t*-statistics. Portfolios are rebalanced every year. The estimations are justified in the presence of both heteroscedasticity and autocorrelation of unknown forms according to Newey and West (1987).

9.4 Conclusions

In this study, we find that British small stocks, on average, display poor performance compared to big stocks in the last two decades. The US, however, shows a slight positive *SMB* over the same period, but the *t*-test results illustrate that *SMB* is not statistically different from zero. Similar results can also be found in Dimson and Marsh (1999) and Ang and Chen (2005). Consistent with the majority of the literature on the value premium, there is a statistically significant value premium in the UK stock market. The return spread between high and low book-to-market decile portfolios (HML_d) is over 10% annually. This *HML* is also pronounced, and comparable with the US results.

We compare Amihud's absolute illiquidity measure and our relative liquidity measure in this study. According to Amihud's measure, small stocks are illiquid where illiquid stocks, on average, show negative returns over time, and liquid stocks have high positive expected returns. The return difference between liquid and illiquid is over 22% annually; however, as expected, Amihud's measure is highly correlated with stock size. Our relative liquidity measure produces little correlation with stock size and any other pervasive risk factors. Nevertheless, the return spread between liquid and illiquid decile portfolios is still striking, at 18% annually.

Cross-sectional analysis shows that there is no size anomaly in the UK from 1991 to 2004, because the CAPM can efficiently explain the excess return of small over big stocks. There is, however, a pronounced value anomaly within this period. The CAPM fails to explain this return difference. A liquidity-augmented CAPM can successfully explain the observed value anomaly.

Finally, the ability of liquidity in explaining value premium is robust to the financial distress factor and a number of macroeconomic variables. The liquidity premium of liquid over illiquid stocks is statistically significant even after being adjusted by the *SMB*, *HML*, distress factor and momentum. Some natural questions arise: what are the underlying risk factors that are responsible for this pronounced liquidity premium? Is liquidity a systematic risk? Is there any connection between liquidity and beta? An unreported result shows that the beta of the most liquid (illiquid) decile portfolio is 1.36 (0.90), and the Wald test highly rejects the equality of these two betas.

The interaction between liquidity and other stock characteristics remains an unexplored area in empirical finance. In modern finance, some of the observed return regularities cannot be explained by the rational asset pricing model – for example, short-term momentum, the cross-sectional difference related to volatility, etc. This could be the result of the incompetence of the models themselves. The success of liquidity in explaining value anomaly in this study gives us much more momentum to pursue further research in this area, and suggests that liquidity is an important component in forecasting expected returns.

References

Agarwal, V. and Taffler, R. (2005). Does the financial distress factor drive the momentum anomaly? Seminar at Cass Business School.

Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time series effects. *Journal of Financial Markets*, 5:31–56.

Amihud, Y. and Mendelson, H. (1986). Asset pricing and the bid–ask spread. *Journal of Financial Economics*, 17:223–249.

Ang, A. and Chen, J. (2005). CAPM over the long run: 1926–2001. *Journal of Empirical Finance*, forthcoming; available at http://www.columbia.edu/~aa610/

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics*, 9:3–18.

Brennan, M. J. and Subrahmanyam, A. (1996). Market microstructure and asset pricing: on the compensation for illiquidity in stock returns. *Journal of Financial Economics*, 41:441–464.

Campbell, J. Y., Grossman, S. J. and Wang, J. (1993). Trading volume and serial correlation in the stock returns. *Quarterly Journal of Economics*, 107:905–939.

Chen, N. and Zhang, F. (1998). Risk and return of value stocks. *Journal of Business*, 71:501–535.

Chordia, T., Roll, R. and Subrahmanyam, A. (2000). Commonality in liquidity, *Journal of Financial Economics*, 56:3–28.

Chordia, T., Subrahmanyam, A. and V. Anshuman, R. (2001). Trading activity and expected stock returns. *Journal of Financial Economics*, 59:3–32.

Dimson, E. and Marsh, P. R. (1999). Murphy's law and market anomalies. *Journal of Portfolio Management*, 25(2):53–69.

Dimson, E., Marsh, P. R. and Staunton, M. (2003a). Global evidence on equity risk premium. *Journal of Applied Corporate Finance*, 15:27–38.

Dimson, E., Nagel, S. and Quigley, G.. (2003b). Capturing the value premium in the UK. *Financial Analysts Journal*, 59(6):36–45.

Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. *Journal of Finance*, 47:427–466.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33:3–56.

Fama, E. F. and French, K. R. (1995). Size and book-to-market factors in earning and returns. *Journal of Finance*, 50(1):131–155.

Fama, E. F. and French, K. R. (1996). Multifactor explanations of asset-pricing anomalies. *Journal of Finance*, 51:55–84.

Fama, E. F. and French, K. R. (1998). Value versus growth: the international evidence. *Journal of Finance*, 53(6):1975–1999.

Fama, E. F. and French, K. R. (2006). The value premium and the CAPM. *Journal of Finance*, 61(5):2163–2185.

Hasbrouck, J. and Seppi, D. J. (2001). Common factors in prices, order flows, and liquidity. *Journal of Financial Economics*, 59:383–411.

Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: implications for stock market efficiency. *Journal of Finance*, 48:65–91.

Kyle, A. (1985). Continuous auctions and insider trading. *Econometrica*, 53:1315–1335.

Lee, C. and Swaminathan, B. (2000). Price momentum and trading volume. *Journal of Finance*, 55(5):2017–2069.

Lo, A. and Wang, J. (2000). Trading volume: definitions, data analysis, and implications of portfolio theory. *The Review of Financial Studies*, 13:257–300.

Newey, W. and West, K. (1987). A simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55:703–708.

Pastor, L. and Stambaugh, R. (2003). Liquidity risk and expected stock returns. *Journal of Political Economy*, 111:642–685.

Petkova, R. and Zhang, L. (2005). Is value riskier than growth? *Journal of Financial Economics*, 78:187–202.

Saretto, A. A. (2004). Predicting and pricing the probability of default. Working Paper, UCLA (available at http://www.personal.anderson.ucla.edu/alessio.saretto/default.pdf).

Silber, W. L. (1975). Thinness in capital markets: the case of the Tel Aviv Stock Exchange. *Journal of Financial and Quantitative Analysis*, 10:129–142.

Zhang, L. (2005). Value premium. *Journal of Finance*, LX(1): 67–103.

10 The information horizon – optimal holding period, strategy aggression and model combination in a multi-horizon framework

Edward Fishwick

10.1 The information coefficient and information decay

The purpose of this chapter is to investigate links between the optimal holding period of a strategy, the aggression of the strategy (which can be related to riskiness in some circumstances), and the risk–return benefits of strategy/model combination.

Whilst this problem will be addressed in a forecasting context, and may be more about active investment than risk management, it has, nevertheless, interesting risk management dimensions. To understand why is to appreciate that the risk dimension of certain strategies changes a great deal with the holding period involved; value investment over short periods will have quite different risk dimensions than value investing over intermediate periods.

IC_1 denotes the cross-sectional information coefficient between model information (forecasts) formed at zero, and returns over some discrete interval of time from zero (i.e. now to one – that is, one step ahead). This is the period ahead IC, and is the conventional meaning of the term.

Now, more generally let IC_t denote the cross-sectional information coefficient over the discrete period $(t-1)$ to (t) for model information formed at time zero. This is *not* the same as the multi-period (i.e. 'cumulative') IC from zero to (t), except in the case of t equals one.

Since the economic value of financial forecast information is almost certainly subject to time decay, in general:

$$IC_1 > IC_t \qquad \forall\ t > 1 \tag{10.1}$$

Thus IC_t decays as the horizon increases.

The functional form of this decay is unknown. Motivated by a desire for realism and tractability, and following Grinold and Kahn (1999), we assume decay of the form:

$$IC_t = IC_1 e^{-bt} \tag{10.2}$$

The parameter b represents the rate of decay with respect to time.

Given values for IC_t the information decay rate (b) can be estimated empirically using ordinary least squares(OLS) in the logs:

$$Ln(IC_t) = Ln(IC_1) - bt \tag{10.3}$$

Re-arranging equation (10.3) and setting IC_1/IC_t equal to 0.5, we find the half-life of a model – i.e. the period over which half of a model's value is used up:

$$Ln(2)/b = t_{\text{half-life}} \tag{10.4}$$

The concept of half-life is useful for understanding the meaning of specific values of the time decay coefficient b. Table 10.1 shows the relationship over some likely values of b.

The multi-horizon 'cumulative' IC to any horizon from zero to (t) is ambiguous in its conception; it could mean the IC added over t periods, or it could mean the IC of t-period cumulative log-returns; we shall adopt the latter notion. We shall approximate this below; it is defined entirely by the one-step-ahead IC_1 and the rate of information decay b, and is given by:

$$IC_{0\to t} = IC_1(1/t)^{0.5}(1 - e^{-bt})/(1 - e^{-b}) \tag{10.5}$$

In equation (10.5), $IC_{0\to t}$ is the IC over the period zero to t. Once the initial IC_1 and the decay rate (b) are known, 'cumulative' forecasting power to any horizon is known by construction in this framework. The division by the square root of t reflects the fact that whilst the standard deviation of cumulative returns increases with the square root of t, the covariance of cumulative returns with $S(0)$, the model information, increases with t.

Let $PIC_{t,s(0)}$ denote the partial IC_t w.r.t. model information at zero – $S(0)$, and define this by:

$$PIC_{t,s(0)} = [IC_t - \Phi S(0)_t IC_1](1 - (\Phi S(0)_t)^2)^{-1} \tag{10.6}$$

where $\Phi S(0)_t$ denotes the correlation between model information (forecasts or scores) at time zero and t. Note that 1C1 is the IC of information formed at zero over the interval zero to one *and equals* the IC of information formed at $t-1$ over the interval $t-1$ to t. This term $PIC_{t,ss(0)}$ corresponds to the regression coefficient of returns from $t-1$ to t in a regression of $S(0)$ on returns from $t-1$ to t and returns from 0 to 1.

Table 10.1 The decay parameter (b) and implied strategy half-life

Time decay parameter (b)	Approximate half-life (months)
0.04	17
0.08	9
0.12	6
0.16	4
0.35	2

$PIC_{t.ss(0)}$ is the explanatory power of model information formed at zero over the period $t-1$ to t that is not due simply to the fact that model information at t is correlated with model information at zero. Setting $PIC_{t,s(0)}$ equal to zero would imply that the only reason that a model is able to forecast more than one period ahead is because new model information is correlated with the information at zero. If we make this assumption, then using equation (10.6) in equation (10.2), gives:

$$\Phi S(0)_t = e^{-bt} \tag{10.7}$$

Thus, *given the assumption* earlier, the decay term e^{-bt} has an interpretation as the correlation between model information (forecasts/scores) at zero and at horizon t.

So far we have discussed a single model. In many problems of interest we have multiple models or forecasters all forecasting the same asset return. If we combine models in composites, the partial ICs in an N model problem are given by:

$$P = [\Omega]^{-1}.B \tag{10.8}$$

where P is a vector – length N of partial ICs $(1, \ldots, N)$, $\{\Omega\}^{-1}$ is the inverse of the $N \times N$ correlation matrix Ω of model forecast correlations, and B is a vector – length N of univariate IC_1s $(1, \ldots, N)$ – i.e. univariate one-step-ahead ICs.

The one-step-ahead IC_1 of the composite model ($IC_{1,comp}$) is then given by:

$$IC_{1,comp} = (P'.B)^{0.5} \tag{10.9}$$

The multi-horizon partial ICs are given by:

$$MHP = [\Omega]^{-1}E.B \tag{10.10}$$

where MHP is a vector (length N) of partial ICs for the n models to some horizon t, and E is an $N \times N$ diagonal matrix (the decay matrix), whose diagonal elements are given by:

$$(1/t)^{0.5}(1-e^{-b_m t})/(1-e^{-b_m}) \tag{10.11}$$

where b_m is the information decay beta specific to model m ($m = 1, \ldots, N$). The horizon t of MHP is determined by the horizon of the decay matrix E.

Finally, the composite IC to any horizon $t - IC_{t,comp}$ is given by

$$IC_{t,comp} = (B'.E.[\Omega]^{-1}.EB)^{0.5} \tag{10.12}$$

10.2 Returns and information decay in the single model case

Conventionally, in a single period framework, and assuming linear conditional expectations functions, expected active returns – i.e. relative returns – are described by:

$$E[R] = IC_1 S(0)\sigma \tag{10.13}$$

where $S(0)$, is the zero mean unit standard deviation model score, and σ is the standard deviation of return (if measurement is monthly, σ is a monthly figure, etc.). Thus expected return equals volatility × score × IC. (The use of equation (10.13) is widespread in practice, and is motivated – knowingly or otherwise – by the fact that equation (10.13), given normality, corresponds to the conditional mean of the return, given the model score.)

Equation (10.13) is almost always used in a single period context. Therefore, in the present context we should re-write it as:

$$E[R_{0\to 1}]_0 = IC_1 S(0)\sigma_{1M} \tag{10.14}$$

where the left-hand variable is the expectation formed at zero of return from zero to one. We assume in equation (10.14), and maintain this assumption going forward, with no loss of generality, that a discrete (unit) increment of time equals 1 month.

Equation (10.5) gives us the IC to any horizon. Therefore, using equation (10.5) in equation (10.14) gives us the return to any horizon:

$$E[R_{0\to t}]_0 = IC_1(1/t)^{0.5}(1-e^{-bt})/(1-e^{-b})S(0)\sigma_{1M}(t)^{0.5} \tag{10.15a}$$

and thus

$$E[R_{0\to t}]_0 = IC_1(1-e^{-bt})/(1-e^{-b})S(0)\sigma_{1M} \tag{10.15b}$$

Equation (10.15b) says that the return to some horizon t is given by the initial score $S(0)$, the IC to the horizon, which is given by IC_1 and the decay rate of information b, and the return volatility over 1 month, the unit of time assumed here.

Since equation (10.15) applies to a varying horizon, and returns to differing horizons are not directly comparable, it is useful to modify this equation to produce annualized returns. It is necessary in this context to capture the impact of transactions costs, since annualizing returns generated at various horizons must imply differing transactions levels. Furthermore, to understand the impact of varying $S(0)$ on the optimal horizon, it is necessary to incorporate transactions costs into the analysis.

We therefore re-write equation (10.15) in annualized after cost form as:

$$E[AR_{0\to t}]_0 = IC_1\left(\frac{1-e^{-bt}}{1-e^{-b}}\right)S(0)\sigma_{1M}(T/t) - C(T/t) \tag{10.16}$$

where C is two-way transactions costs, and T is the number of increments of discrete time in 1 year. For example, if t is measured in months, T equals 12 etc. (NB: equation (10.16) is derived by multiplying equation (10.15b) by (T/t) to produce annualized returns before cost, and then subtracting the annualized cost term $C(T/t)$. This clearly assumes that returns are additive, which we justify by assuming that σ is measured in logs.)

Equation (10.16) says that the expected return generated by a forecast depends on six things:

1. One period IC (IC_1)
2. Information decay rate (b)
3. Strategy aggression (S)

4. Transactions costs (C)
5. Security volatility (σ)
6. Holding period (t).

Given values for (1) to (6), it is possible to find the optimal horizon (or holding period) t for a given forecast or model(s) by maximizing equation (10.16) subject to the parameter values (1) to (6). We are thus able to address two questions. First, given some model, what is the optimal investment horizon? Secondly, what are expected returns at the optimal horizon? In the conventional single period framework this is not possible.

The above implies that the (optimal) investment horizon is not a matter of choice, but is an intrinsic property of any given model(s) or process. That is, there is some optimal level of turnover, equivalent to the holding period (determined by combination of the information decay rates with certain other strategy parameters), which maximizes annualized return. Varying the operational horizon from this optimum will reduce after-cost return.

We wish to understand how the optimal holding period varies with information decay rate, transactions costs and strategy aggression. We assume initial default values for the variables (1) to (5), and examine the relationship between changes in these parameters, and the optimal horizon and return. The default values are as follows:

1. One period IC_1 – default value = 0.04
2. Information decay rate – default value = 0.12
3. Strategy aggression (i.e. score) – default value = 1.0
4. Transactions costs – default value = 0.80
5. Security volatility – default value = 7.20.

The default values are not unrealistic. A monthly IC of 0.04 is within the range of observed ICs in practice (and not inconsistent with practitioner experience). An information decay rate of 0.12 implies an information half-life of a little under 6 months (and is not inconsistent with practitioner experience). The strategy aggression of score equal to 1.0 implies an expected return of +1 standard deviation to the investment horizon. The security volatility of 7.2 equates to an annualized value of 24.95, which is approximately the median volatility of securities in the Russel 1000 over the period 89.12 to 97.12.

Table 10.2 summarizes the direction of the relevant relationships. Essentially, it shows the sign on the left-hand column variables as they impact holding period and return.

The results in Table 10.2 show – on a partial basis – that the faster information decays, the shorter the optimal holding period and the lower the return. Increasing strategy

Table 10.2 The impact of varying strategy parameters on the optimal horizon

	Holding period	**Return**
Information decay	Negative relationship	Negative relationship
Strategy aggression	Negative relationship	Positive relationship
Transactions costs	Positive relationship	Negative relationship

aggression decreases the optimal holding period and increases the return. Increasing transaction costs increases the optimal holding period and decreases return.

Tables 10.3–10.5 give some precise parametric results, from the application of the NAG routine nag_opt_simplex to equation (10.16). We maximize annualized return subject to the parameter values above, and thus solve for the optimal horizon (t).

First, we vary the information decay rate, allowing model half-life to vary from 17 months down to just 3 months. The optimal holding period varies from around 14 months in the case of slow information down to 8 months in the case of the faster-decaying information.

Note that in the case of the slow information decay (long half-life), we hold stocks for less than the half-life of the model. In the case of the fast information decay (short

Table 10.3 The rate of information decay and optimal horizon

Decay (Half-life in months)	Optimal holding (months)	Return (p.a.)
0.04 (17)	14.0	2.06
0.08 (9)	10.7	1.57
0.12 (6)	9.5	1.23
0.16 (4)	8.8	0.96
0.20 (3)	8.4	0.75

Table 10.4 Strategy aggression (score) and the optimal horizon

Score	Optimal holding (months)	Return (p.a.)
0.4	24.5	0.07
0.6	14.8	0.36
1.0	9.6	1.15
1.4	7.5	2.06
1.8	6.3	3.04

Table 10.5 Transactions costs and optimal horizon

Costs	Optimal holding (months)	Return (p.a.)
0.2	3.9	2.28
0.5	6.5	1.58
0.8	9.6	1.15
1.2	13.4	0.72
1.8	21.1	0.29

half-life), we hold stocks for nearly three times the half-life. This pays for higher turnover relative to the half-life.

Secondly, we vary strategy aggression. We allow the model 'score' to vary between 0.4 (low aggression) and 1.8 (high aggression). The optimal holding period varies between around 2 years for the lowest aggression strategy and 6 months for the high aggression case. Notice that – ignoring for the moment the issue of deadweight – an optimal holding period of 2 years for the low aggression case implies an active turnover of 50% per year, whilst in the high aggression case an optimal holding period of 6 months implies an annual turnover of 200%.

Finally in this section we assess the impact of varying transactions costs on the optimal holding period. We allow two-way transactions costs to range from a low of 0.2% to a high of 1.8%. Variations in transactions costs over this range have a major impact on the optimal holding period.

Note in Table 10.5 that the impact of transactions costs on return is not constant across the range of costs in question. Reducing transactions costs by 60 basis points from 1.8 to 1.2 increases return by only 43 basis points. However, reducing transactions costs by 60 basis points from 0.8 to 0.2 increases return by 113 basis points. This is because as transactions costs fall, the optimal horizon shortens and thus turnover increases. In the case of a fall from 1.8 to 1.2, turnover is always below 100%. In the case of a fall from 0.8 to 0.2, turnover is above 150%. Clearly, reducing cost has a higher impact on return in the latter case.

This relationship is important for aggressive, short-horizon, high-turnover strategies that incur high transactions costs.

10.3 Model combination

Equation (10.8) can be used to define how models are combined in a conventional single horizon framework. Since the optimal horizon will vary with strategy parameters, such as aggression and transactions costs, the weighting of models in a composite must also depend on those strategy parameters.

Thus, increasing the aggression of a strategy will modify the optimal model mix (assume the b_m varies across models). Likewise changes in the cost of transactions – including potentially tax – will cause the optimal model weights to change.

The annualized return to any horizon t in the N sub-model case is given by:

$$[R_{0\to t}]_0 = S'.[\Omega]^{-1}EB.\sigma_{1M}(T/t) - C(T/t) \tag{10.17}$$

where, as in equation (10.8), B is a vector – length N – of univariate IC_1s, Ω is the $N \times N$ correlation matrix of model forecasts, E (the decay matrix) is a diagonal matrix whose elements are given by equation (10.11), $\{S\}$ (the score vector) is a $1 \times N$ row vector with elements $S(0)_m$, $(m = 1, \ldots, N)$ and $\sigma_{1M}(T/t)$ is a scalar representing the horizon adjusted volatility. As in equation (10.16), the product $C(T/t)$ is a scalar representing annualized transactions costs at model horizon t.

Equation (10.17) is simply the N model version of equation (10.16) and can be maximized in the same way – solving for expected return at the optimal horizon by finding the horizon t that maximizes after cost annualized return, given a set of strategy parameters

(our earlier variables (1) to (6)). Clearly, however, the added complication here is that as the horizon t varies with the strategy parameters, the implied weights on the N models will vary given varying b_m.

In general, increasing strategy aggression will tend to increase the weight on high-decay beta models (short half-life), whilst decreasing aggression will tend to move the optimal model combination toward low-decay beta (long half-life) models.

10.4 Information decay in models

We focus on some apocryphal funds VALUE1 (value), VALUE2 (cash flow), and MOMENT (momentum) and a composite COMPOS (combined). We took the time series of the quintile spread of the real time *ex ante* forecasts generated by these models for the 7-year period 89.12 to 97.12 in the top 1000 US universe. We warn readers that in what follows the values are typical rather than actual.

For convenience, we transformed the quintile spreads into ICs. We measured the non-overlapping ICs for horizons +1 month through +18 months, and fitted log linear OLS as specified in equation (10.3) to this multi-horizon data. We constrained the regression to pass through the point IC(1) – i.e. we assume that information decays, *from* the one period ahead IC – and estimated the following decay factors. The numbers reported are similar qualitatively to the sorts of numbers one might get in practice (see Table 10.6).

The 'value' type models (VALUE1 and VALUE2) have long horizons. The 'momentum' model (MOMENT) has much shorter horizon. The composite (COMPOS) lies between the two. In the short term, the 'momentum' information is more valuable than the 'value' type information. In the longer term, however, the 'value' type information remains valuable, whilst the momentum-based forecasts become worthless quite quickly.

The realized multi-horizon ICs of these models were used to estimate the information decay of the models. Following our earlier assumptions, we can examine the *underlying forecasts* and derive a separate estimate of time decay, utilizing the assumption used to derive equation (10.7).

For four non-overlapping 24-month periods (89 to 97) we measured the 24 months of partial correlation between cross-sectional model forecasts. We thus have four non-overlapping estimates of the value of $\Phi S(0)_t$ for t equals 24 – the cross-sectional correlation of model forecasts 24 months apart.

Re-arranging equation (10.7) gives:

$$b = \ln(1/\Phi S(0)_t)t^{-1} \tag{10.18}$$

Table 10.6 Information decay in models, from realized multi-horizon ICs

Model	IC_1	Decay beta	Half-life (months)
VALUE1	0.024	0.02	34.6
VALUE2	0.016	0.03	27.5
MOMENT	0.030	0.19	3.6
COMPOS	0.037	0.17	4.2

Table 10.7 Information decay in models from long horizon signal correlation

	$\Phi S(0)_{24}$	Implied decay beta	Implied half-life
VALUE1	0.27	0.05	13.07
VALUE2	0.21	0.06	10.99
MOMENT	0.07	0.12	5.91
COMPOS	0.12	0.09	7.54

Thus, the estimate of cross-sectional forecast correlation gives an estimate of time decay. These estimates for our models are shown in Table 10.7.

The rank ordering of the decay betas from the two estimation methodologies (Tables 10.6 and 10.7) is the same. The VALUE1 and VALUE2 models have longer horizons, MOMENT has a short horizon. The COMPOS model is, by construction, in between. The actual decay betas – and therefore half-lives – are materially different, however. Two potential reasons are worthy of consideration:

1. Both estimates may be subject to estimation error – contamination by noise
2. The assumption used to derive the estimate in equation (10.7) (that $PIC_{t,S(0)}$ equals zero) may be invalid – the values for the decay betas in Table 10.6 may be correct, and the value of $PIC_{t,S(0)}$ in equation (10.6) may be non-zero. That is, the forecasting power of old information may differ from that implied by the simple correlation between old and new information.

If $PIC_{t,S(0)} \neq$ zero, the implication from the results in Tables 10.6 and 10.7 is as follows. For the value type sub-composites (VALUE1 and VALUE2), $PIC_{t,S(0)}$ is positive. Thus, on a partial basis, old information is positively correlated with longer horizon return. For MOMENT and the COMPOS, PIC_1 is negative. Thus old information is negatively correlated with longer horizon return on a partial basis.

We note without further exploration that the above is consistent with a situation in which market pricing 'overreacts' to momentum-type information, and 'underreacts' to value-type information.

Given that both methods of estimation may in part be correct, we work from here with a simple combination (rounded average) of the results from the two methods of estimation of b_m (see Table 10.8).

It may be possible to improve these estimates using more sophisticated techniques or more extensive analysis. However, given the probable time-varying nature of the decay

Table 10.8 Composite estimates of information decay in models

	IC	Decay beta	Half-life (months)
VALUE1	0.020	0.03	23.0
VALUE2	0.015	0.04	17.0
MOMENT	0.032	0.17	4.0
COMPOS	0.039	0.14	5.0

beta, the likely levels of noise, the limited data, and variations in model specification over time, we suspect that the estimates in Table 10.8 may be hard to better.

10.5 Models – optimal horizon, aggression and model combination

Finding the optimal horizon in the single model case involves maximizing equation (10.16). That is, the optimal horizon is the one that maximizes the conditional asset return. Using the model parameters from Table 10.8 (IC and decay beta), for the COMPOS model, a range of optimal horizons dependent on aggression and transactions costs is as shown in Table 10.9.

Finding the optimal horizon using the sub-composites constitutes an N model case, and therefore involves maximizing equation (10.17). As noted previously, this will also define sub-model 'weights' at the optimal horizon.

In order to solve equation (10.17), we require the correlation matrix Ω. Using the time series of the quintile spreads generated by the sub-composites, we calculated the following (rounded) correlations: VALUE1/VALUE2 = 0.45; VALUE1/MOMENT = −0.05; VALUE2/MOMENT = 0.15. These seem reasonably intuitive – the value/growth correlation is slightly negative, the value/cash flow price measure is high and positive, and the growth/cash flow price measure is small but positive – and we therefore use them at this stage to form Ω.

If we assume transactions costs of 1%, and the parameter values for the decay matrix from Table 10.8, maximization of equation (10.17) gives estimates of optimal horizons and weights as shown in Table 10.10.

Table 10.9 COMPOS – optimal horizon in months

Transactions costs	0.6	0.8	1.0	1.2
Score = 0.6	12	15	19	26
Score = 1.0	8	10	12	14
Score = 1.4	6	7	9	10
Score = 1.8	5	6	7	8

Table 10.10 Optimal horizon and implied weights for sub-composites

	Optimal horizon (months)	% VALUE1	% VALUE2	% MOMENT
Score = 0.42	24	56	10	34
Score = 0.60	16	52	9	39
Score = 1.00	10	48	7	45
Score = 1.40	8	45	7	48
Score = 1.80	7	44	5	51
Score = 2.20	6	43	4	53

Table 10.11 Map of average portfolio ranking onto score

Average decile ranking	Average score
0.5	2.11
1.0	1.78
1.5	1.41
2.0	1.16
2.5	0.97
3.0	0.80
3.5	0.65
4.0	0.50
4.5	0.35

To put these scores in context, consider this mapping of *average* portfolio decile ranking onto average score (Table 10.11).

Thus an average decile ranking of around 4.3 implies an optimal horizon of 24 months which, ignoring deadweight, implies a turnover of 50% per annum, with sub-composite weights VALUE1 = 56%, MOMENT = 34%, and VALUE2 = 10%.

A more aggressive portfolio – say an average decile ranking of 2.4 – implies an optimal horizon of 10 months, and thus turnover of 120%. Now, however, the optimal weight on MOMENT is 45%, whilst that on VALUE1 and VALUE2 has fallen to 48% and 7% respectively.

As strategy aggression rises, the average holding time falls with the optimal horizon. This means that the slow decaying information (VALUE1 and VALUE2) becomes relatively less valuable, and thus these slow decay models are de-weighted in favour of the faster information (MOMENT).

One considers a standard mean-variance efficient frontier. In conventional analysis, the returns model used in generating the efficient frontier will have constant weights on its sub-components across the domain of the feasible set – i.e. there is a common model generating returns. The fact is that portfolio aggression increases significantly between two points, so that the distribution changes over different points in mean-variance space; from the perspective of a mean-variance investor, this does not matter, although it does matter in reality.

It is clear in the present framework, however, that as aggression (score) increases the optimal horizon decreases. Thus, as tracking error increases, the implied turnover of the strategy is increasing. Therefore, as we move out along the efficient frontier, the information horizon is shortening as strategy aggression increases. We demonstrated in Section 10.3, though, that as the optimal horizon changes, optimal model weights also vary. Specifically, as we move out along the efficient frontier we would expect to de-emphasize slow decay information, and emphasize faster decaying information that is initially more powerful.

Assume a model that generates return forecasts from a set of sub-components with constant weights, and assume that risk varies across the sub-components. There are then two possibilities. First, it may be that the constant weight model is never optimal with respect to the information horizon (there is no reason why it should be, so long as $N > 2$). Secondly, since optimal horizon varies along the x axis of a mean-risk diagram, it may

be that at a single point on the efficient frontier the constant weight model is optimal (for $N = 2$ this must be true; for $N > 2$ this may occur by chance). Assume in mean-risk space that, at co-ordinates risk equals 2.5, score equals 2.0, the constant weight model is optimal. That is, at the information horizon implied at that point the constant weights are equal to the optimal weights at that point.

Above risk = 2.5, however, the constant weight model gives too much weight to slow decay information. Thus there is some sub-model combination that dominates. Likewise, below risk = 2.5 the constant weight model gives too much weight to fast decay information and again, there must be some sub-model combination that dominates the constant weight model.

It is highly unlikely that the constant weight model is optimal. The *magnitude* of the *deviation* between the constant weight and optimized to horizon approach will depend on actual model parameter values. In general, the wider the range of b_m and IC_1 values, the greater the potential benefits. (This in turn suggests that in model research and development, the creation and selection of models with a wide array of b_m values may also be a worthwhile goal.)

Reference

Grinold, R. and Kahn, R. (1999). *Active Portfolio Management*. Chicago, IL: Probus.

11 Optimal forecasting horizon for skilled investors

Stephen Satchell and Oliver Williams

11.1 Introduction

The concept of forecasting skill can be modelled in traditional asset pricing frameworks by defining the joint distribution between an investor's forecast of future returns and actual returns themselves. Literature such as Grinold and Kahn (1999) defines the Information Coefficient (IC) to be the correlation between these forecast and actual returns which provides a basis for parameterizing pricing models so that effects of skill can be investigated from various points of view such as equilibrium price determination and comparative statics.

The elementary concept of IC as a single correlation can be readily extended to a multi-period setting in either discrete or continuous time. In such a situation IC can be defined in various different ways according to the timing of forecasts and the time horizon of the respective realized returns. A natural way to approach this is to consider correlation between a set of forecasts made *simultaneously* at time zero but in respect of returns realized over a *series* of contiguous future time periods.

Multi-period IC therefore introduces the possibility that an investor can have differential forecasting skill depending on time horizon. It is quite intuitively plausible, for instance, that a forecaster who excels on a near-term basis might nevertheless deteriorate when predicting longer-term returns (and in the very distant future limit it is unreasonable to expect IC to be significantly non-zero). Similarly, one forecaster may display different profiles of skill over time when predicting returns on different assets: returns on 'value' stocks may be more successfully forecast on a longer time horizon when some reversion to fundamental valuation has occurred, whereas returns on 'growth' stocks may tend to be highly auto-correlated over certain periods, making short time horizon forecasts quite accurate but longer-term forecasts hit-or-miss.

Fishwick (2007) presents a detailed analysis of the problem, and an optimal time horizon solution for forecasting skill. However, the solution is numerical rather than exact, and a closed-form solution with clear corresponding comparative statics would be highly desirable.

This chapter addresses that particular question. We choose a power functional form for time decay of information rather than the exponential process assumed by Fishwick. This extends Fishwick (2007) by rewriting many of the calculations in continuous rather than discrete time. This will include the discrete time as a special case, but allows us to use calculus with respect to optimal time. The theory is presented in Section 11.2.

We also allow for multiple sources of model forecast deterioration by allowing the variance of the forecasted returns to be a function of the forecast horizon. This seems to

us to be eminently sensible. It is well known that forecasts typically become more volatile as the forecast horizon increases. As we demonstrate, this does not complicate the results unduly.

One specific advantage of making time continuous is that it allows us to derive an explicit formula for the optimal time horizon of the model as a function of the following parameters: the initial IC, the score, the deterioration rate of the forecasted ICs, the growth of the forecasts' volatility, and a cost factor. These calculations are presented in Section 11.3.

In Section 11.4, we recast the multi-model optimal horizon problem to simultaneously solve the optimal portfolio and optimal time horizon problem.

In Section 11.5, we suggest an alternative formulation of the multi-model problem which is, we believe, closer to Grinold's original dictum of alpha equals IC times score times volatility (Grinold, 1994). Whichever formulation one adopts, however, it is conceptually possible to optimize both model weights and time simultaneously.

Finally, we present conclusions in Section 11.6.

11.2 Analysis of the single model problem

11.2.1 Preliminary definitions and symbols

Assumption 1: *$S(0)$ is the conditioning information known at time 0. It is assumed that*:

$$S(0) \sim N(0, 1)$$

Assumption 2: *$R(r)$ is the instantaneous forecast rate of return (alpha) from the model at time r.*

$$R(r) \sim N(0, \sigma^2(r))$$

Assumption 3: *S(0) and R(r) follow a bivariate normal stochastic process.*

Assumption 4: *IC(r) is the Information Coefficient between R(r) and* S(0)*, i.e.*

$$IC(r) = Corr(R(r), S(0))$$
$$= \frac{Cov(R(r), S(0))}{\sigma(r)}$$

11.2.2 Definition of the continuous rate of return

From the above assumptions it follows that the continuous rate of return

$$R[0, t] = \int_0^t R(r)dr$$

has the following properties:

$$R[0,t] \sim N\left(0, \int_0^t \sigma^2(r)dr\right)$$

If $R(r)$ is serially uncorrelated, in the general case, we define a covariance function

$$Cov(R(s), R(r)) = C(s, r)$$

where the variance is equal to

$$\int_0^t \int_0^t C(s, r)dsdr$$

From the formula for the conditional expectation of a bivariate normal, and Assumption 1, we can compute the beta, $\beta(t)$, as

$$\begin{aligned}\beta(t) &= E(R[0,t]/S(0)|S(0)) \\ &= \frac{Cov(R[0,t], S(0))}{Var(S(0))} \\ &= Cov\left(\int_0^t R(r)dr, S(0)\right) \\ &= \int_0^t Cov(R(r), S(0)dr) \\ &= \int_0^t \sigma(r)IC(r)dr\end{aligned}$$

Defining the *gross expected return* to the model, conditional on information per unit time as

$$\frac{\beta(t)}{t}S(0)$$

and defining a *total cost* function $C(t)$ and a cost per unit time as

$$\frac{C(t)}{t}$$

we can then define the *net conditional expected rate of return, W(t)*, as

$$\begin{aligned}W(t) &= \frac{\beta(t)}{t}S(0) - \frac{C(t)}{t} \\ &= \frac{\beta(t)S(0) - C(t)}{t} \\ &= \frac{H(t)}{t}\end{aligned}$$

11.2.3 Optimal time horizon

In this section we characterize the properties of the optimal time horizon t^*. Following Fishwick (2007), t^* is chosen as the value that maximizes $W(t)$.

In a world of no alpha uncertainty or in a world of risk neutrality, maximizing $W(t)$ is entirely sensible. Grinold refers to an alpha-associated risk that he calls active risk. We shall delay discussion of whether to bring risk into the problem until we consider the multi-model problem. Introduction of active risk is highly problematic, as Grinold acknowledges. There are difficult practical issues of how to measure alpha volatility. There are also interesting theoretical issues relating optimal forecast horizons to the forecaster's expected utility (this subject is considered by Williams (2007)).

Theorem 1: The value of t that maximizes $W(t)$, t^*, satisfies the following conditions, assuming $H(t) > 0$:

$$H'(t^*)t^* - H(t^*) = 0$$

$$S(0)\beta''(t^*) < C''(t^*)$$

Proof: see Appendix A.

Theorem 2: (Comparative statics)

$$\frac{\partial t^*}{\partial S(0)} > 0 \qquad if \ \frac{\beta(t^*)}{t^*} < \beta'(t^*)$$

$$\frac{\partial t^*}{\partial S(0)} < 0 \qquad if \ \frac{\beta(t^*)}{t^*} > \beta'(t^*)$$

Proof: see Appendix A.

11.2.4 Discussion of theorems

Theorem 2 tells us that the optimal time t^* increases with the information $S(0)$ if the average value of $\frac{\beta(t^*)}{t^*}$ is less than the marginal value $\beta'(t^*)$, and decreases with an increase in $S(0)$ if $\frac{\beta(t^*)}{t^*}$ is greater than $\beta'(t^*)$.

$\beta(t^*)$ is the volatility-weighted cumulative IC. In the case that $\sigma(r) = \sigma$ for all r, and $C(t)$ is constant and equal to C, i.e. the assumptions in Fishwick, Theorem 1 requires that $\beta(t)$ has a second derivative less than zero. This means that the cumulative IC must have a negative second derivative (concavity), which is equivalent to $IC(r)$ being decreasing in r. Note that in this case:

$$\beta(t) = \sigma \int_0^t IC(r)dr$$

and

$$\text{the cumulative IC} = \frac{\beta(t)}{\sigma} \tag{11.1}$$

Theorem 2 simply says that if the average cumulative IC is less than the marginal IC, an increase in $S(0)$ (aggression) leads to an increase in t^*, whilst if average cumulative IC is greater than the marginal IC, an increase in aggression leads to a decrease in t^*. For the case in Fishwick (equation 10.2), the average is less than the marginal so we would expect an increase in strategy aggression to decrease t^*. This is what Fishwick finds in Table 10.2.

11.2.5 Modelling IC decay

In Theorems 1 and 2 we do not specify the volatility function or the IC function. We shall now consider examples of the above. We define the class of initial value dependent IC functions as follows:

Definition 1: $IC(r)$ is an initial value dependent IC function if it has a representation

$$IC(r) = IC(0)g(r) \tag{11.2}$$

for $g(r)$ some positive function with $\frac{\partial g(r)}{\partial r} < 0$

Remark 1: The Fishwick function (in discrete time) is

$$IC(r) = IC(1)\exp(-b(r-1)) \tag{11.3}$$

for r a positive integer. This can be generalized in continuous time to

$$IC(r) = IC(0)\exp(-br) \tag{11.4}$$

We now see that such a specification as equation (11.2) leads to substantial advantages in the case of constant volatility,

$$\sigma(r) = \sigma$$

In particular,

$$\beta(t) = IC(0)\int_0^t g(r)dr$$

$$\beta'(t) = IC(0)g(t)$$

and if

$$C(t) = C(\text{constant costs})$$

Theorem 1 becomes

$$\sigma IC(0)S(0)g(t^*)t^* = \sigma IC(0)S(0)\int_o^{t^*} g(r)dr - C \tag{11.5}$$

and

$$g'(t^*) < 0 \qquad \text{if } S(0) > 0$$

11.3 Closed-form solutions

The above equation (11.5) allows us to get closed-form solutions for optimal time t^*.

Suppose the decay function is

$$g(r) = r^{-\alpha} \qquad 0 < \alpha < 1 \tag{11.6}$$

then for $S(0) > 0$ and $C > 0$ we find

$$t^{*(1-\alpha)} = \frac{t^{*(1-\alpha)}}{1-\alpha} - \frac{C}{\sigma S(0) IC(0)}$$

$$\left(\frac{\alpha}{1-a}\right) t^{*(1-\alpha)} = \frac{C}{\sigma S(0) IC(0)}$$

$$t^* = \left(\frac{C(1-\alpha)}{\sigma\alpha S(0) IC(0)}\right)^{\frac{1}{1-\alpha}} \tag{11.7}$$

The above function gives us comparable results to the Fishwick function. However, the Fishwick function is more realistic as $r^{-\alpha}$ behaves more erratically as r increases relative to $exp(-br)$. Also, as informally illustrated in Figure 11.1, $0 < \exp(-br) < 1 \forall r > 0$ and b positive whilst $r^{-\alpha}$ only has this property for $r > 1, 0 < \alpha < 1$. For both sets of values both functions satisfy the appropriate second-order conditions for Theorem 1 and 2. The only advantage of equation (11.6) is equation (11.7), a closed-form solution. This is not obtainable with equation (11.4). However in the case of equation (11.4), the continuous time generalization of the Fishwick function, a solution for t^* can be obtained in terms of the Lambert W function (this solution is presented in Appendix B).

It is worth noting that if we take logs of equation (11.7) and differentiate, we can compute elasticities:

$$\frac{\partial \ln(t^*)}{\partial \ln(C)} = \frac{1}{1-\alpha} > 1 \tag{a}$$

$$\frac{\partial \ln(t^*)}{\partial \ln(S(0))} = \frac{-1}{1-\alpha} < -1 \tag{b}$$

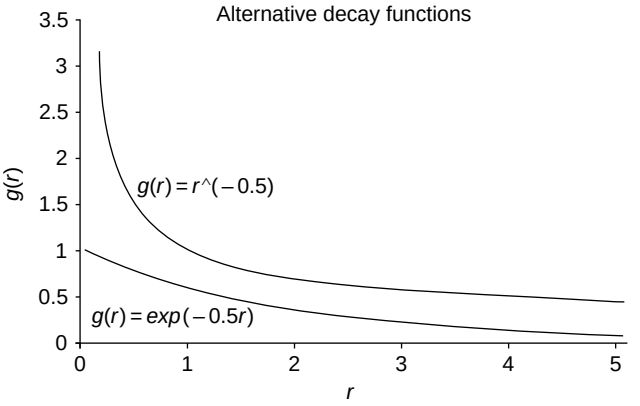


Figure 11.1 Qualitative comparison of alternative decay functions $g(r)$.

$$\frac{\partial \ln(t^*)}{\partial \ln(IC(0))} = \frac{-1}{1-\alpha} < -1 \tag{c}$$

$$\frac{\partial \ln(t^*)}{\partial \ln(\sigma)} = \frac{-1}{1-\alpha} < -1 \tag{d}$$

and

$$\frac{\partial \ln(t^*)}{\partial \alpha} = \left(\frac{1}{1-\alpha}\right)\left(\ln(t^*) - \frac{1}{\alpha(1-\alpha)}\right) \gtrless 0 \tag{e}$$

All the above are trivial except (e). Since there is potential disagreement about what the sign of (e) should be, we provide a proof in Appendix A. These are summarized in Table 11.1, alongside some similar results obtained for the continuous-time Fishwick function.

Table 11.1 Elasticities and sensitivities of optimal time with respect to various parameters.

	$g(r)=r^{-\alpha}$ (Elasticity)	$g(r)=\exp(-br)$ (Derivative)*
C	>1	>0
$S(0)$	<-1	<0
$IC(0)$	<-1	<0
σ	<-1	<0
α	$\gtrless 0$	
b		<0

* derivations appear in Appendix B.

The above have the following interpretations:

(a) The optimal time increases with costs; its elasticity with regard to costs is greater than one (elastic)
(b) The optimal time decreases with aggression; its elasticity with regard to aggression is greater than one (elastic)
(c) The optimal time decreases with volatility; its elasticity with regard to volatility is greater than one (elastic)
(d) The optimal time decreases with $IC(0)$, its elasticity with regard to $IC(0)$ is greater than one (elastic)
(e) An increase in the decay rate (α) may increase or decrease the optimal time.

Comparing our results with those of Fishwick (Table 10.2), we see that both models have the following in common:

1. Increasing costs increases the optimal time
2. Increasing aggression decreases the optimal time.

However, they differ on information decay, as in our case $g(r) = r^{-\alpha}$ and increasing information decay (α) may increase or decrease the optimal time, whilst in the Fishwick case increasing information decay b decreases the optimal time. We analyze this further next.

The result in (e) merits further discussion. It says that the optimal time increases with α 'information decay' if $t^* > \exp\left(\frac{1}{\alpha(1-\alpha)}\right)$. Notice that since $0 < \alpha < 1$, $\exp\left(\frac{1}{\alpha(1-\alpha)}\right)$ has a minimum of $\exp(4)$ when $\alpha = \frac{1}{2}$ and otherwise is larger. Thus, for values of α near 0 or 1, we would expect to find t^* less than this number, which could be very large. Speaking heuristically, as long as t^* is large we will get sensible results.

We present a theorem that relates t^* to the shape of $g(r)$. Let $g(r) = g(r, \theta)$ where θ is some shape parameter. We find the following result:

Theorem 3:

$$sign\left[\frac{\partial t^*}{\partial \theta}\right] = sign\left[\left(\frac{\partial g(t^*, \theta)}{\partial \theta} t^*\right) - \int_0^t \partial \frac{g(r, \theta)}{\partial \theta} dr\right]$$

Proof: See Appendix A.

Discussion of Theorem 3

Theorem 3 explains how the shape of the decay function influences the impact of information decay on the optimal time. One can verify that Theorem 3 for $g(r) = \exp(-br)$ will imply that $sign\left[\frac{\partial t^*}{\partial b}\right]$ is negative. Whilst we have not verified that $sign\left[\frac{\partial t^*}{\partial \theta}\right]$ is positive or negative via Theorem 3, it perhaps suggests that there are problems in interpreting θ as a decay parameter. This again relates to the problems for $g(r) = r^{-\theta}$ when $0 < r < 1$. One can show that $\frac{\partial g}{\partial r} = -r^{-\theta}\ln(r)$ which is negative for $r > 1$, zero when $r = 1$ and *positive*

when $0 < r < 1$. Thus θ is a decay parameter only when $r > 1$ and the perverse behaviour of $g(r)$ for $0 < r < 1$ explains why we can get the wrong sign.

11.3.1 Adapting the closed-form solution to increasing forecast horizon volatility

We now assume that $\sigma(r)$ is changing in r; this means that 2-year forecasts have different volatility from 1-year forecasts. The term $\sigma(r)$ is the standard deviation of the forecast of returns at time r. Plausibly, one could imagine scenarios where this quantity could be increasing, decreasing or constant in r.

To combine with our previous results, we now assume that

$$\sigma(r) = \sigma r^{\theta} \tag{11.8}$$

Such models exist in the literature and are known as CEV (constant elasticity of volatility) models, but where r is a state variable (often price). They are usually employed in the pricing of options and the modelling of term structure.

Here, r is the investment horizon. Under the assumptions of equations (11.6) and (11.8):

$$\begin{aligned}\beta(t) &= IC(0)\sigma \int_0^t r^{\theta-\alpha} dr \quad 0 < \alpha < 1 \\ &= IC(0)\sigma \frac{t^{\theta-\alpha+1}}{\theta-\alpha+1} \\ \beta'(t) &= IC(0)\sigma t^{\theta-\alpha}\end{aligned}$$

Recalling from Section 11.2.2:

$$\begin{aligned}H(t) &= \beta(t)S(0) - C(t) \\ &= IC(0)\sigma \frac{t^{\theta-\alpha+1}}{\theta-\alpha+1} S(0) - C(t)\end{aligned}$$

Making the assumption of time-invariant costs C gives:

$$H'(t) = S(0)IC(0)\sigma t^{\theta-\alpha}$$

and so applying Theorem 1 leads to

$$H'(t^*)t^* - H(t^*) = 0$$

$$S(0)IC(0)\sigma t^{*\theta-\alpha+1} - S(0)IC(0)\sigma \frac{t^{*\theta-\alpha+1}}{\theta-\alpha+1} + C = 0$$

$$t^{*\theta-\alpha+1}\left(\frac{\theta-\alpha}{\theta-\alpha+1}\right) = \frac{-C}{\sigma S(0)IC(0)}$$

Given $S(0) > 0$ and $IC(0) > 0$ we require that $\theta < \alpha$ for this equation to have a positive solution, thus

$$t^* = \left[\frac{C(1-(\alpha-\theta))}{(\alpha-\theta)\sigma S(0) IC(0)}\right]^{\frac{1}{1-(\alpha-\theta)}} \qquad \text{for } 0 < \alpha - \theta < 1$$

As before, we find that, rewriting the results in Section 11.3,

$$\frac{\partial \ln(t^*)}{\partial \ln(C)} = \frac{1}{1-(\alpha-\theta)} > 1$$

$$\frac{\partial \ln(t^*)}{\partial \ln(S(0))} = \frac{-1}{1-(\alpha-\theta)} < -1$$

$$\frac{\partial \ln(t^*)}{\partial \ln(IC(0))} = \frac{-1}{1-(\alpha-\theta)} < -1$$

$$\frac{\partial \ln(t^*)}{\partial \ln(\sigma)} = \frac{-1}{1-(\alpha-\theta)} < -1$$

$$\frac{\partial \ln(t^*)}{\partial(\alpha-\theta)} \gtreqless 0$$

so

$$\frac{\partial \ln(t^*)}{\partial \alpha} \gtreqless 0 \quad \text{and} \quad \frac{\partial \ln(t^*)}{\partial \theta} \gtreqless 0$$

The qualitative conclusions are as before, except that an increase in the pattern of forecast volatility can increase or decrease the optimal time horizon. The results above do not depend upon $\theta > 0$, only $\theta < \alpha$. Thus our results will apply for decreasing patterns of volatility as well as increasing patterns.

11.4 Multi-model horizon framework

When an investor has access to forecasts from multiple different models, an intriguing question is, how should these forecasts be combined in order to determine an overall single prediction of future returns? This problem can be approached in alternative ways.

11.4.1 Partial information coefficient

A fundamental concept in Fishwick's method of multi-model optimization is that of Partial Information Coefficient (PIC). For the case of k multiple models ($i = 1 \ldots k$) the *PIC(t)* for model i represents the correlation between actual returns in period t and the period t forecast of model i after controlling for the period t forecasts of all other models $\neq i$. By standard results in multivariate statistics, this is given by:

$$\underline{\mathbf{PIC}} = \mathbf{COR}^{-1}\underline{\mathbf{IC}} \tag{11.9}$$

An equivalent approach to *PIC* is to consider the regression paradigm. Suppose we have k models, each with its own scalar score $S^{(i)}$ and information coefficient $IC^{(i)}$. From assumptions 3 and 4, it follows that the forecast of model i will be given by $\sigma IC^{(i)} S^{(i)}$. Recalling that $S^{(i)}$ is defined to have unit standard deviation, we can interpret the quantity $\sigma IC^{(i)}$ as a regression beta between actual returns R and the model i score. Note that this is the beta which would be computed in a regression model in which $S^{(i)}$ is the *only* explanatory variable. Again applying standard results from linear regression, it is clear that if all model scores $S^{(i)}, i \in \{1 \ldots k\}$ are *orthogonal* then the $\sigma IC^{(i)}$ beta would remain unchanged if scores of any other models were added as additional regressors. Indeed, one might expect that adding these further regressors would tend to improve goodness of fit.

However, in the case of *non-orthogonal* models, where $\mathbf{COR} \neq \mathbf{I}$ (the $(k \times k)$ identity matrix), the betas associated with each individual model will depend on the particular set of models used as explanatory variables. To state this symbolically in regression terms, let $\beta^{(i)}$ be the beta associated with model i in the multiple regression which incorporates ***all*** k models as explanatory variables (as will become apparent, this will be defined as $PIC^{(i)}\sigma$). Then:

$$\begin{bmatrix} \beta^{(1)} \\ \beta^{(2)} \\ \vdots \\ \beta^{(k)} \end{bmatrix} = Cov(\mathbf{S}, \mathbf{S})^{-1} Cov(\mathbf{S}, R)$$

$$= Cov(\mathbf{S}, \mathbf{S})^{-1} Cov\left(\begin{bmatrix} S^{(1)} \\ S^{(2)} \\ \vdots \\ S^{(k)} \end{bmatrix}, R \right)$$

$$= (\mathbf{COR})^{-1} E\left(\begin{bmatrix} S^{(1)} \\ S^{(2)} \\ \vdots \\ S^{(k)} \end{bmatrix} R \right)$$

$$= (\mathbf{COR})^{-1} E \begin{bmatrix} S^{(1)} S^{(1)} IC^{(1)} \sigma \\ S^{(2)} S^{(2)} IC^{(2)} \sigma \\ \vdots \\ S^{(k)} S^{(k)} IC^{(k)} \sigma \end{bmatrix}$$

$$= (\mathbf{COR})^{-1} \begin{bmatrix} IC^{(1)} \sigma \\ IC^{(2)} \sigma \\ \vdots \\ IC^{(k)} \sigma \end{bmatrix}$$

Hence defining $PIC^{(i)}\sigma = \beta^{(i)}$ and dividing both sides above by σ reduces this to

$$\begin{bmatrix} PIC^{(1)} \\ PIC^{(2)} \\ \vdots \\ PIC^{(k)} \end{bmatrix} = (\mathbf{COR})^{-1} \begin{bmatrix} IC^{(1)} \\ IC^{(2)} \\ \vdots \\ IC^{(k)} \end{bmatrix}$$

which is equivalent to (11.9) as given by Fishwick.

Leading on from the definition of *PIC*, Fishwick proposes that when multiple models are available their forecasts should be weighted according to the $PIC^{(i)}$ values. These would seem to be a natural choice if one considers the $PIC^{(i)}$s to be simply regression betas divided by returns standard deviation.

In a multi-period setting we will be interested in computing returns forecasts for a specific time horizon t. In this case, following on from the decay concepts introduced in Section 11.2.5, we will be interested in computing $PIC^{(i)}$s consistent with a certain profile $IC^{(i)}(t)$. This can be easily incorporated into calculations by dealing with the vector $\mathbf{V}\underline{\mathbf{IC}}(1)$ instead of $\underline{\mathbf{IC}}(1)$ by itself, where $\mathbf{V}$ is a $(k \times k)$ diagonal matrix which premultiplies each initial period $IC^{(i)}$ by a factor in order to give cumulative *IC* for the period from zero to t.

Specifically, denoting this cumulative *IC* by $IC(0, t)$ it is clear that

$$\begin{aligned} IC(0,t) &\equiv Corr\left(\sum_{j=1}^{t} R(j), S\right) \\ &= \frac{Cov\left(\sum_{j=1}^{t} R(j), S\right)}{\sqrt{Var\left(\sum_{j=1}^{t} R(j)\right) Var(S)}} \\ &= \frac{\sum_{j=1}^{t} \sigma(j) IC(j)}{\sqrt{\sum_{j=1}^{t} \sigma^2(j).1}} \\ &= \frac{\sigma \sum_{j=1}^{t} IC(j)}{\sigma\sqrt{t}} \\ &= \sqrt{\frac{1}{t}} \sum_{j=1}^{t} IC(j) \end{aligned}$$

(where we make the simplifying constant variance assumption $\sigma(j) = \sigma$) and therefore in the discrete time exponential decay case $\mathbf{V}$ will be a $(k \times k)$ diagonal matrix with elements given by:

$$V_{ij} = \sqrt{\frac{1}{t}} \frac{1 - \exp(-b_i t)}{1 - \exp(-b_i)} \qquad \text{if } i = j$$

$$V_{ij} = 0 \quad \text{if } i \neq j$$

since from equation (11.3) we have

$$\sqrt{\frac{1}{t}}\sum_{r=1}^{t} IC(j)$$
$$= \sqrt{\frac{1}{t}} IC(1) \sum_{r=1}^{t} \exp(-b(r-1))$$
$$= \sqrt{\frac{1}{t}} IC(1) \frac{1-\exp(-bt)}{1-\exp(-b)}$$

Hence Fishwick lays out an equation for expected returns with k models in his Equation 10.17 reproduced below (subject to a minor notational change: Fishwick's denotes our matrix **V** as **E**):

$$E(AR[0,t]) = \underline{\mathbf{S}}'(\mathbf{COR})^{-1}\mathbf{V}\underline{\mathbf{IC}}(1)\frac{\sigma_{1m}}{t} - \frac{C}{t} \tag{11.10}$$

where:

$E(AR[0,t])$ is the expected average net conditional rate of return from 0 to t, $\underline{\mathbf{S}}$ is a $(k \times 1)$ vector of model scores, **COR** is the $(k \times k)$ matrix of correlations between model scores, and $\underline{\mathbf{IC}}(1)$ is a $(k \times 1)$ vector incorporating the k initial ICs of the separate models.

It is further assumed that $\sigma_{ij} = \sigma_{1m}$ for $j = 1$, i.e. the standard deviations of the alphas of each model are the same and they are constant throughout time per unit time.

(Note that for clarity of notation matrices are printed in boldface and vectors are bold underlined.)

11.4.2 Discussion of equation (11.10)

We shall consider some minor changes to the above definition. Later we shall present an alternative version of the multi-model net conditional expected rate of return.

This definition weighs each model by the score of the model; it is by no means clear that this is appropriate.

An alternative approach is to let **S** be a $(k \times k)$ diagonal matrix with the score $S_i(0)$ on the i^{th} diagonal, zero off the diagonal and employ a weight vector $\underline{\mathbf{w}}$, such that $\underline{\mathbf{w}}'\underline{\mathbf{e}} = 1$.

Likewise in the above definition C becomes $\underline{\mathbf{C}}$, a $(k \times 1)$ vector of costs. We can recover equation (11.10) above by letting $\underline{\mathbf{C}} = C\underline{\mathbf{e}}$ where $\underline{\mathbf{e}}$ is a $(k \times 1)$ vector of ones. With this amendment to the structure, we are now able to incorporate the realistic feature of differential costs between models which is not possible in the original Fishwick framework.

Assuming a covariance matrix of $\underline{\mathbf{AR}}[0,t]$ to be $\boldsymbol{\Sigma}(t)$ where $\boldsymbol{\Sigma}(t)$ is a $(k \times k)$ matrix depending on t, we have:

$$\underline{\mathbf{w}}'\underline{\Pi} = E(\underline{\mathbf{w}}'\underline{\mathbf{AR}}[0,t]) = \underline{\mathbf{w}}'\mathbf{S}'\mathbf{V}\underline{\mathbf{IC}}(1)\frac{\sigma_{1m}}{t} - \frac{\underline{\mathbf{w}}'\underline{\mathbf{C}}}{t} \tag{11.11}$$

$$Var(\underline{\mathbf{w}}'\underline{\mathbf{AR}}[0,t]) = \underline{\mathbf{w}}'\boldsymbol{\Sigma}(t)\underline{\mathbf{w}}$$

in which element i in the vector of model-by-model net returns $\underline{\mathbf{\Pi}}$ is evidently obtained as the product $S_i(0)V_{ii}IC_i(1)\frac{\sigma_{1m}}{t}$ minus $\frac{C_i}{t}$, i.e. the product of score, average cumulative IC to period t and volatility σ_{1m} minus average model-specific cost.

The issue now arises as to whether we can simultaneously optimize expected utility to choose w and t^* jointly. For simplicity we write $\mathbf{\Sigma} = \mathbf{\Sigma}(t)$.

Notice that if we minimize the variance of $\underline{\mathbf{w}}'\underline{\mathbf{AR}}[0, t]$ so that $\underline{\mathbf{w}}'\underline{\mathbf{\Pi}} = \mu$, $\underline{\mathbf{w}}'\underline{\mathbf{e}} = 1$, we simply recover the usual MV frontier parametrized on t, i.e. for any given t there is a mean-variance frontier.

We write the optimization as

$$\min u = \frac{1}{2}\underline{\mathbf{w}}'\mathbf{\Sigma}\underline{\mathbf{w}} - \boldsymbol{\lambda}'(\mathbf{\Phi}\underline{\mathbf{w}} - \underline{\mathbf{G}}) \tag{11.12}$$

where

$$\mathbf{\Phi} = (\underline{\mathbf{\Pi}}, \underline{\mathbf{e}})'$$

and

$$\underline{\mathbf{G}} = (\mu, 1)'$$

the (2×1) vector of constants. Straightforward optimization leads to

$$\hat{\underline{\mathbf{w}}} = \mathbf{\Sigma}^{-1}\mathbf{\Phi}'(\mathbf{\Phi}\mathbf{\Sigma}^{-1}\mathbf{\Phi}')^{-1}\underline{\mathbf{G}} \tag{11.13}$$

and the minimized variance is given by

$$\delta(t) = \underline{\mathbf{G}}'(\mathbf{\Phi}\mathbf{\Sigma}^{-1}\mathbf{\Phi}')^{-1}\underline{\mathbf{G}}$$

Finally we minimize $\delta(t)$ with respect to t. Up till now both $\mathbf{\Phi}$ (the forecasts) and $\mathbf{\Sigma}$ (the covariance matrix) could be time-dependent. We shall assume now that $\mathbf{\Sigma}$ is fixed and does not depend on t. Although this is not necessary, it leads to some simplification. Then we have

$$\frac{\partial u}{\partial t} = -\underline{\mathbf{G}}'(\mathbf{\Phi}\mathbf{\Sigma}^{-1}\mathbf{\Phi}')^{-1}\left[\frac{\partial \mathbf{\Phi}}{\partial t}\mathbf{\Sigma}^{-1}\mathbf{\Phi}' + \mathbf{\Phi}\mathbf{\Sigma}^{-1}\frac{\partial \mathbf{\Phi}'}{\partial t}\right](\mathbf{\Phi}\mathbf{\Sigma}^{-1}\mathbf{\Phi}')^{-1}\underline{\mathbf{G}} = 0$$

The optimal time t^* is the value of t that minimizes u for a given return $\underline{\mathbf{\Pi}}$.

It is probably more sensible to choose the value of t that maximizes expected utility in which case we need to minimize functions of equation (11.12) with respect to t. Whether the problem has a well-defined maximum or not is an unsolved question.

An alternative, and more general, solution is not to consider the optimal time but to maximize expected utility for a portfolio $\underline{\mathbf{w}}$ and a vector of times $(t_1, \ldots, t_R)$.

Grinold defines score (Grinold, 1994: p. 11) as: 'a standardised measure that shows how strongly you feel about a particular stock at a particular time'.

The above definition, and subsequent discussion in Grinold, suggests that one interpretation of the score is information known at time zero that we can condition our future forecasts (alphas, or equivalently $R(r)$) upon. In a multi-model problem with multiple

forecasts, it is perfectly natural to consider a single feeling for the stock/index that the multiple models are forecasting. The benefit of this interpretation is that equation (11.11) simplifies to:

$$\underline{\mathbf{w}}'\underline{\mathbf{\Pi}} = S(0)\underline{\mathbf{w}}'\mathrm{V}\underline{\mathbf{IC}}(1)\frac{\sigma_{1m}}{t} - \frac{\underline{\mathbf{w}}'\underline{\mathbf{C}}}{t} \tag{11.14}$$

11.5 An alternative formulation of the multi-model problem

11.5.1 Review of Grinold's formula

To consider the multi-model problem, we need to clarify what is the appropriate generalization of Grinold's formula, namely alpha equals score times volatility times *IC*. It is our contention that Grinold's formula derives from assumptions 1, 2, 3 and 4 in Section 11.2.1

Thus:

$$\begin{aligned} &Alpha(r) \\ &\quad = E(R(r)|S(0)) \\ &\quad = \frac{Cov(R(r), S(0))}{Var(S(0))}(S(0) - E(S(0)) + E(R(r)) \\ &\quad = S(0)\sigma(r)IC(r) \end{aligned} \tag{11.15}$$

To generalize this to k models we need to consider how equation (11.15) changes if we allow $\underline{\mathbf{R}}(r)$ to be a $(k \times 1)$ vector. The interpretation of $\underline{\mathbf{R}}(r)$ is the vector of the instantaneous expected rates of return from the k different models at time r. Furthermore, $S(0)$ is the scalar score. In this situation, for example, we have one current piece of information about a single company $S(0)$ and we have k different models that forecast the company. It follows immediately that $Alpha(\mathrm{r})$ is now a $(k \times 1)$ vector and that

$$\begin{aligned} &Alpha(r) \\ &\quad = Cov(\underline{\mathbf{R}}(r), S(0))S(0) \end{aligned} \tag{11.16}$$

For equation (11.16) to be true, we have assumed:

Assumption 5: *a single score*, $S(0)$, $E(S(0)) = 0$, $Var(S(0)) = 1$

Assumption 6: $E(\underline{\mathbf{R}}(r)) = 0$

Assumption 7: $\underline{\mathbf{R}}(r)$ *and* $S(0)$ *have a* $(k+1)$*-dimensional Gaussian stochastic process.*

We can simplify equation (11.16) further by noting that, for a $(k \times k)$ diagonal matrix $\mathbf{D}(r)$ which has the k standard deviations $\sigma_i(r),\ i = 1..k$ of the k forecasts $R_i(r)$, the following is true:

$$\begin{aligned} Cov(\mathbf{R}(r), S(0)) &= \mathbf{D}(r) Corr(\underline{\mathbf{R}}(r), S(0)) \\ &= \mathbf{D}(r)\underline{\mathbf{IC}}(r) \end{aligned}$$

where $\underline{\mathbf{IC}}(r)$ is the $(k \times 1)$ vector of the ICs of the k models for a horizon r.

If we now want to compute the cumulative alpha $[0, t]$:

$$\begin{aligned} &Alpha[0, t] \\ &= E\left(\int_0^t \underline{\mathbf{R}}(r)dr | S(0)\right) \\ &= \left(\int_0^t \mathbf{D}(r)\underline{\mathbf{IC}}(r)dr\right) S(0) \end{aligned}$$

If we assumed constant forecast volatility we have:

$$\mathbf{D}(r) = \mathbf{D} \forall r$$

and then

$$\begin{aligned} &Alpha[0, t] \\ &= \mathbf{D}\left(\int_0^t \underline{\mathbf{IC}}(r)dr\right) S(0) \end{aligned}$$

If, further,

$$\underline{\mathbf{IC}}(r) = \mathbf{G}(r)\underline{\mathbf{IC}}(0)$$

for $\mathbf{G}(t)$ an appropriate diagonal matrix and $\underline{\mathbf{IC}}(0)$ a $(k \times 1)$ vector,

$$\begin{aligned} &Alpha[0, t] \\ &= \mathbf{D}\left(\int_0^t \mathbf{G}(r)dr\right) \underline{\mathbf{IC}}(0) S(0) \end{aligned}$$

11.5.2 A comparison of different multi-model formulations

To see why our approach is fundamentally different from Fishwick's equation 10.17, let us write $\underline{\mathbf{R}}(r)$ as $\mathbf{x}$, a $(k \times 1)$ vector, and $S(0)$, a scalar, as y.

Define the covariance matrix of $(\mathbf{x}, y)$ as:

$$\begin{bmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{x}y} \\ \Sigma_{y\mathbf{x}} & \Sigma_{yy} \end{bmatrix}$$

Note that by Assumption 5, $(Var\ S(0) = 1)$.

The regression coefficients of y on $\mathbf{x}$ is $\beta_{y\mathbf{x}} = \mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{\Sigma}_{\mathbf{x}y}$ whilst the regression coefficients of $\mathbf{x}$ on y is $\beta_{\mathbf{x}y} = \mathbf{\Sigma}_{y\mathbf{x}}$. (Since $\mathbf{\Sigma}_{yy} = \mathbf{I}$).

Using the fact that $\mathbf{D}$, the $(k \times k)$ diagonal matrix of standard deviations, is assumed to be

$$\mathbf{D} = \sigma_{1m}\mathbf{I}_k$$

where $\mathbf{I}_k$ is the $(k \times k)$ identity matrix, we see that

$$\begin{aligned}\Sigma_{\mathbf{xx}} &= \mathbf{D}(\mathbf{COR})\mathbf{D} \\ &= \sigma_{1m}^2\mathbf{COR} \\ \Sigma_{\mathbf{x}y} &= \sigma_{1m}IC\end{aligned}$$

Thus

$$\begin{aligned}\beta_{y\mathbf{x}} &= (\sigma_{1m}^2\mathbf{COR})^{-1}\sigma_{1m}IC \\ &= \frac{1}{\sigma_{1m}}(\mathbf{COR})^{-1}IC\end{aligned}$$

whilst

$$\beta_{\mathbf{x}y} = \sigma_{1m}IC$$

Fishwick's equation 10.17 employs $\beta_{y\mathbf{x}}$ whilst our proposed solution, equation (11.15), uses $\beta_{\mathbf{x}y}$. We argue, however, that $\beta_{\mathbf{x}y}$ is the correct formula to use as we should be conditioning $\underline{\mathbf{R}}(r)$ on $S(0)$; we should not be conditioning $S(0)$ on $\underline{\mathbf{R}}(r)$.

Determining what is the correct interpretation hinges on what $S(0)$, the score, actually is. We need to see if the inventors of the concept can help us. Grinold and Kahn (1995: 218) claim that the score is 'the raw forecast that contains the active manager's information in raw form: an earnings estimate, buy or sell recommendation, etc.'. The raw forecast can come in a variety of units and scales, and is not directly a forecast of exceptional return. It is clear from this definition that we could have many models and many raw forecasts. Our analysis in Sections 11.4 and 11.5 deals with the case of many models, one score. Results can be calculated for the multiple models multiple score case. We omit the details, but the methodology is essentially similar.

11.6 Conclusions

The contributions of this chapter are the following:

- We rebuild the model in Fishwick's paper in continuous time.

- The reformulation described in Section 11.3 allows us to get a closed-form expression for the optimal time.
- To achieve the closed-form expression we need to replace exponential decay of information by power decay of information. The cost of this is that there is, in one particular case, some ambiguity about the parametric representation of information decay.
- The methodology does allow us to confirm the results in Fishwick's paper and also allows us to investigate the impact of changing patterns of forecast volatility.
- In Sections 11.4 and 11.5 we consider problems to do with combining model forecasts. Here we are not convinced that Grinold's methodology, or Fishwick's methodology, are the only ways to proceed.
- We present several interpretations of the problem, based on generalizations of the famous dictum, alpha is IC times score times volatility. These lead us away from the existing methodologies.
- We present an argument which gives a framework to compute optimal model weights and optimal times. We leave as an open question an analysis of the conditions on the model forecast which guarantee a well-defined solution to this difficult problem.

Appendix A

Proof of Theorem 1

Let

$$w(t) = \frac{H(t)}{t}$$

then

$$\frac{\partial w(t)}{\partial t} = 0 \qquad iff \;\; tH'(t) = H(t)$$

also

$$\begin{aligned}\frac{\partial^2 w(t)}{\partial t^2} &= \frac{\partial}{\partial t}\left[\frac{tH'(t) - H(t)}{t^2}\right] \\ &= \frac{t^2[tH''(t) + H'(t) - H'(t)] - 2t[tH'(t) - H(t)]}{t^4} \\ &= \frac{H''(t)}{t} \qquad \text{when } t = t^*\end{aligned}$$

So

$$\frac{\partial^2 w(t^*)}{\partial t^2} < 0 \qquad \text{iff } H''(t) < 0$$

$$\text{or } \beta''(t)S(0) < C''(t)$$

Proof of Theorem 2

Totally differentiating the equation in Theorem 1 with respect to $S(0)$ (and dropping the * superscript on (t) we obtain

$$t\beta'(t) - \beta(t) + (S(0)t\beta''(t) - tC''(t))\frac{\partial t}{\partial S(0)} = 0$$

Since $S(0)t\beta''(t) - C''(t)t < 0$ from Theorem 1, we see that $\frac{\partial t}{\partial S(0)}$ is positive if $t\beta'(t) - \beta(t)$ is positive, as claimed.

Proof of Theorem 3

$$g(t^*, \theta)t^* = \int_0^{t^*} g(r, \theta)dr + C'$$

where θ is some parameter governing the shape of $g(t^*)$ and

$$C' = \frac{C}{IC(0)S(0)\sigma}$$

Differentiating implicitly,

$$\frac{\partial g}{\partial t^*}t^*\frac{\partial t^*}{\partial \theta} + \frac{\partial g}{\partial \theta}t^* + g\frac{\partial t^*}{\partial \theta} = \int_0^{t^*} \frac{\partial g}{\partial \theta}dr + g\frac{\partial t^*}{\partial \theta}$$

and therefore

$$\frac{\partial t^*}{\partial \theta}\left(t^*\frac{\partial g}{\partial t^*}\right) = \int_0^{t^*} \frac{\partial g(r, \theta)}{\partial \theta}dr - \frac{\partial g(t^*, \theta)}{\partial \theta}t^*$$

or

$$\frac{\partial t^*}{\partial \theta} = \frac{\int_0^t \frac{\partial g(r,\theta)}{\partial \theta}dr - \frac{\partial g(t^*,\theta)}{\partial \theta}t^*}{t^*\frac{\partial g}{\partial t^*}}$$

By assumption, $\frac{\partial g}{\partial t^*} < 0$ so that

$$\frac{\partial t^*}{\partial \theta} > 0 \qquad \text{if } \int_0^t \frac{\partial g(r, \theta)}{\partial \theta}dr < \frac{\partial g(t^*, \theta)}{\partial \theta}t^*$$

and

$$\frac{\partial t^*}{\partial \theta} < 0 \qquad \text{if } \int_0^t \frac{\partial g(r, \theta)}{\partial \theta}dr > \frac{\partial g(t^*, \theta)}{\partial \theta}t^*$$

Proof of Section 11.3 elasticity (e)

Let

$$C^* = \frac{C}{\sigma S(0) IC(0)}$$

so

$$t^* = \left[\frac{C^*(1-\alpha)}{\alpha}\right]^{\frac{1}{1-\alpha}}$$

then

$$\ln(t^*) = \frac{1}{1-\alpha}\ln(C^*) + \frac{1}{(1-\alpha)}\ln(1-\alpha) - \frac{1}{(1-\alpha)}\ln(\alpha)$$

and hence

$$\begin{aligned}\frac{\partial \ln(t^*)}{\partial \alpha} &= \frac{\ln(C^*)}{(1-\alpha)^2} - \frac{1}{(1-\alpha)^2} + \frac{\ln(1-\alpha)}{(1-\alpha)^2} - \frac{\ln(\alpha)}{(1-\alpha)^2} - \frac{1}{\alpha(1-\alpha)} \\ &= \frac{1}{(1-\alpha)^2}\left[\ln(C^*) - 1 + \ln(1-\alpha) - \ln(\alpha) - \frac{(1-\alpha)}{\alpha}\right] \\ &= \frac{1}{(1-\alpha)^2}\left[\ln(C^*) + \ln\frac{(1-\alpha)}{\alpha} - \frac{1}{\alpha}\right] \\ &= \frac{1}{(1-\alpha)}\left[\ln(t^*) - \frac{1}{\alpha(1-\alpha)}\right]\end{aligned}$$

with therefore

$$\frac{\partial \ln(t^*)}{\partial \alpha} > 0 \qquad \text{if } \ln(t^*) > \frac{1}{\alpha(1-\alpha)} \text{ or } t^* > \exp\left(\frac{1}{\alpha(1-\alpha)}\right)$$

and

$$\frac{\partial \ln(t^*)}{\partial \alpha} < 0 \qquad \text{if } \ln(t^*) < \frac{1}{\alpha(1-\alpha)} \text{ or } t^* < \exp\left(\frac{1}{\alpha(1-\alpha)}\right)$$

Appendix B

First, expanding the definitions given in Section 11.2:

$$\begin{aligned}\beta(t) &= \int_0^t \sigma IC(r)dr \\ &= IC(0)\int_0^t \sigma e^{-bt}dr\end{aligned}$$

$$
\begin{aligned}
&= \sigma IC(0)\left[-\frac{1}{b}e^{-br}\right]_0^t \\
&= -\frac{1}{b}\sigma IC(0)[e^{-bt}-1] \\
\beta'(t) &= \sigma IC(t) \\
&= \sigma IC(0)e^{-bt} \\
H(t) &= \beta(t)S(0) - C \\
&= -\frac{1}{b}\sigma IC(0)S(0)[e^{-bt}-1] - C \\
H'(t) &= \beta'(t)S(0) \\
&= \sigma IC(0)S(0)e^{-bt}
\end{aligned}
$$

Then conditions for optimal t^* as per Theorem 1:

$$
\begin{aligned}
&H'(t^*)t^* - H(t^*) = 0 \\
&\sigma IC(0)S(0)e^{-bt^*}t^* + \frac{1}{b}\sigma IC(0)S(0)\left[e^{-bt^*}-1\right] + C = 0 \\
&e^{-bt^*}t^* + \frac{1}{b}\left[e^{-bt^*}-1\right] = \frac{-C}{\sigma IC(0)S(0)} \\
&e^{-bt^*}\left[t^* + \frac{1}{b}\right] = \frac{1}{b} - \frac{C}{\sigma IC(0)S(0)} \qquad (11.17)
\end{aligned}
$$

Making the substitution

$$\theta = -bt^* - 1$$

gives

$$
\begin{aligned}
-bt^* &= \theta + 1 \\
t^* &= -\frac{1}{b}[\theta + 1]
\end{aligned}
$$

and

$$t^* + \frac{1}{b} = -\frac{1}{b}[\theta+1] + \frac{1}{b} = -\frac{\theta}{b}$$

Rewriting equation (11.17):

$$-\frac{\theta}{b}e^{\theta+1} = \frac{1}{b} - \frac{C}{\sigma IC(0)S(0)}$$

$$\theta e^{\theta+1} = \frac{Cb}{\sigma IC(0)S(0)} - 1$$

$$\theta e^{\theta} = e^{-1}\left[\frac{Cb}{\sigma IC(0)S(0)} - 1\right] \tag{11.18}$$

from Corless *et al.* (1996: 332) the solution for θ in equation (11.18) is given by the Lambert W function, i.e.

$$\theta = W\left(e^{-1}\left[\frac{Cb}{\sigma IC(0)S(0)} - 1\right]\right)$$

and so

$$t^* = -\frac{1}{b}\left\{W\left(e^{-1}\left[\frac{Cb}{\sigma IC(0)S(0)} - 1\right]\right) + 1\right\} \tag{11.19}$$

Due to the structure of the Lambert function the following conditions on parameters are required to ensure t^* positive:

$$W\left(e^{-1}\left[\frac{Cb}{\sigma IC(0)S(0)} - 1\right]\right) < -1$$

$$-e^{-1} < e^{-1}\left[\frac{Cb}{\sigma IC(0)S(0)} - 1\right] < 0$$

$$-1 < \frac{Cb}{\sigma IC(0)S(0)} - 1 < 0$$

$$0 < \frac{Cb}{\sigma IC(0)S(0)} < 1$$

For comparative statics the following results are useful. First of all, a standard result is that

$$W'(z) = \frac{1}{(1 + W(z))\exp(W(z))}$$

$$= \frac{W(z)}{z(1 + W(z))} \text{ for } z \neq 0$$

and it is helpful to note from equation (11.19) that

$$t^* = -\frac{1}{b}\left\{W\left(e^{-1}\left[\frac{Cb}{\sigma IC(0)S(0)} - 1\right]\right) + 1\right\}$$

$$-bt^* = W\left(e^{-1}\left[\frac{Cb}{\sigma IC(0)S(0)} - 1\right]\right) + 1$$

$$-bt^* - 1 = W\left(e^{-1}\left[\frac{Cb}{\sigma IC(0)S(0)} - 1\right]\right)$$

hence we have

$$W'\left(e^{-1}\left[\frac{Cb}{\sigma IC(0)S(0)} - 1\right]\right)$$

$$= \frac{W\left(e^{-1}\left[\frac{Cb}{\sigma IC(0)S(0)} - 1\right]\right)}{e^{-1}\left[\frac{Cb}{\sigma IC(0)S(0)} - 1\right]\left(1 + W\left(e^{-1}\left[\frac{Cb}{\sigma IC(0)S(0)} - 1\right]\right)\right)}$$

$$= \frac{-bt^* - 1}{-bt^* e^{-1}\left[\frac{Cb}{\sigma IC(0)S(0)} - 1\right]}$$

$$= \frac{1 + bt^*}{bt^* e^{-1}\left[\frac{Cb}{\sigma IC(0)S(0)} - 1\right]}$$

From this we obtain

$$\frac{\partial t^*}{\partial C} = -\frac{1}{b}\left(\frac{1 + bt^*}{bt^* e^{-1}\left[\frac{Cb}{\sigma IC(0)S(0)} - 1\right]}\right) e^{-1}\left[\frac{b}{\sigma IC(0)S(0)}\right]$$

$$= -\frac{1 + bt^*}{bt^*\left[\frac{Cb}{\sigma IC(0)S(0)} - 1\right]\sigma IC(0)S(0)}$$

$$> 0$$

$$\frac{\partial t^*}{\partial S(0)} = \frac{1}{b}\left(\frac{1 + bt^*}{bt^* e^{-1}\left[\frac{Cb}{\sigma IC(0)S(0)} - 1\right]}\right) e^{-1}\left[\frac{Cb}{\sigma IC(0)[S(0)]^2}\right]$$

$$= \left(\frac{1 + bt^*}{bt^*\left[\frac{Cb}{\sigma IC(0)S(0)} - 1\right]}\right)\left[\frac{C}{\sigma IC(0)[S(0)]^2}\right]$$

$$< 0$$

$$\frac{\partial t^*}{\partial IC(0)} = \frac{1}{b}\left(\frac{1+bt^*}{bt^* e^{-1}\left[\frac{Cb}{\sigma IC(0)S(0)}-1\right]}\right) e^{-1}\left[\frac{Cb}{\sigma S(0)[IC(0)]^2}\right]$$

$$= \left(\frac{1+bt^*}{bt^*\left[\frac{Cb}{\sigma IC(0)S(0)}-1\right]}\right)\left[\frac{C}{\sigma S(0)[IC(0)]^2}\right]$$

$$< 0$$

$$\frac{\partial t^*}{\partial \sigma} = -\frac{1}{b}\left(\frac{1+bt^*}{bt^* e^{-1}\left[\frac{Cb}{\sigma IC(0)S(0)}-1\right]}\right) e^{-1}\left[\frac{Cb}{IC(0)S(0)\sigma^2}\right]$$

$$= \left(\frac{1+bt^*}{bt^*\left[\frac{Cb}{\sigma IC(0)S(0)}-1\right]}\right)\left[\frac{C}{IC(0)S(0)\sigma^2}\right]$$

$$< 0$$

References

Corless, R. M., Gonnet, G. H., Hare, D.E.G. *et al.* (1996). On the Lambert W function. *Advances in Computational Mathematics*, 5:329–359.

Fishwick, E. (2007). Information horizons. *Forecasting Expected Returns*. Oxford: Butterworth-Heinemann.

Grinold, R. C. (1994). Alpha is volatility times IC times score. *Journal of Portfolio Management*, 20:9–16.

Grinold, R. C. (1997). The information horizon. *Journal of Portfolio Management*, Fall (1997):57–67.

Grinold, R. C. and Kahn, R. N. (1999). *Active Portfolio Management*, Second Edition. McGraw-Hill.

Williams, O. J. (2007). Doctoral Dissertation, Faculty of Economics, University of Cambridge.

12 Investments as bets in the binomial asset pricing model

David Johnstone

Abstract

This chapter examines investments within the Cox–Ross–Rubinstein binomial asset pricing model, and describes simple mathematical parallels between long and short positions in the underlying asset and conventional bets made at risk neutral betting odds. The concepts of 'log optimal' investment and 'Kelly betting' are reconciled, and a formula provided for replicating Kelly bets with a portfolio of riskless bonds and derivatives of the underlying asset. At a practical level, the chapter defines a new derivative called a '$1 Kelly bet'. This is an easily implementable betting instrument that enables traders in sportsbetting and other prediction markets to automatically manage an arbitrary sum of capital over repeated bets in the way that exploits their abilities to estimate probabilities to the greatest possible long run monetary effect.

12.1 Introduction

Betting on horse races or football games and investment in stocks are often viewed as essentially the same activity (e.g. Thorp, 1971, 2000; Hausch *et al*., 1981; Asch *et al*., 1982; Ziemba and Hausch, 1985; Thaler and Ziemba, 1988; Pope and Peel, 1989; Hausch and Ziemba, 1990; Golec and Tamarkin, 1991; Shin, 1993; Hausch *et al*., 1994; Woodland and Woodland, 1994; Gray and Gray, 1997; Vaughan Williams and Paton, 1997; Gandar *et al*., 1998; Sauer, 1998; Vaughan Williams 1999, 2005; Ozgit 2005; Smith *et al*., 2006).

> There are many parallels between trading in financial markets and sports wagering. . . . in both settings, investors with heterogeneous beliefs and information seek to profit through trading as uncertainty is resolved over time.
>
> (Levitt 2004: 223)

The purpose of this chapter is to clarify and extend the formal connections between investment and betting within conventional binomial asset pricing models. Any investment position (long or short) in a binary asset can be replicated by a conventional bet made at odds implicit within the market price of that asset (together with the risk-free rate). An interesting aspect of this replication is that the relevant betting odds are 'risk neutral'. That is, they are based on 'risk-neutral' probabilities rather than 'actual' probabilities, and are therefore observable.

Asset pricing theorists have explained that risk-neutral probabilities are not probabilities in a natural sense (e.g. Boyle 1992: 156; Stoll and Whaley, 1993: 203–204). Instead, they are 'pseudo-probabilities' or weights used to find arbitrage-free asset prices via an ingenious shortcut that avoids explicit reference to actual probabilities.

The argument presented below suggests that risk-neutral probabilities are interpretable in the same way as natural ('physical' or 'subjective') probabilities in at least one concrete sense. By buying or selling an asset, the investor makes a bet on the asset going up or down at effective betting odds derived from risk-neutral probabilities. These odds, or their corresponding risk-neutral probabilities, do not represent the market's aggregated subjective beliefs, but they do indicate actual monetary payoffs, the same as betting odds at the racetrack.

Extending the analogy between betting and investment, it is shown that a 'Kelly bet', otherwise known as a 'log optimal' or 'growth optimal' investment, is a simple derivative of the underlying asset and can be replicated with a weighted portfolio of call options, put options and risk-free bonds. The payoff from Kelly betting is a function of the market-implied (risk-neutral) probability and the gambler's subjective probability of the event observed. The only other factor in the payoff function is the risk-free interest rate. Bets with this payoff function can be made in whatever dollar amount the gambler's utility function demands.

To formalize 'Kelly betting' and allow for the possibility that investors might like to gamble just a fraction of their wealth growth-optimally, we introduce the notion of a '$1 Kelly bet'. This is the bet – in effect, the portfolio of bets – that a true Kelly (log utility) gambler would make if his wealth was just $1. Described another way, it is a derivative of the underlying asset that replicates the portfolio that an investor would hold if he had set aside $1 and wanted to invest this amount growth-optimally in the underlying asset.

An investor could possess wealth of $100 and elect to buy twenty '$1 Kelly bets', so as to invest 20% of his wealth in a fund that will grow as quickly as possible, subject to the accuracy of his probability assessments, over the long run. The remainder could be invested more risk aversely, to allow perhaps for the possibility that he is not very good at judging probabilities and hence is likely to lose some or most of the 20% set aside for Kelly betting.

As with other binary assets, it is possible to write options on Kelly bets. These can be valued using conventional binomial pricing models. Depending on the gambler's rationality or ready cash, he might prefer to buy or sell options on Kelly bets rather than the underlying asset (the bets themselves). From the perspective of a true Kelly gambler, there is no difference between investing in options and investing in the underlying asset. Terminal utility is the same either way, and depends simply on the accuracy of the gambler's probability forecasts relative to those of the market.

12.2 Actual versus risk-neutral probabilities

Risk-neutral probabilities, regarded correctly as not 'real' probabilities in the way of either relative frequencies or personal degrees of belief, are often mistaken as the market's beliefs. To see the different roles of actual and risk-neutral probabilities in asset valuation, consider the following non-standard 'decision theoretic' derivation of the price of a call option in a binomial tree (the more standard derivation is shown in Cox and Rubinstein 1985: 172–173; Ingersoll, 1987: 308–309; and Luenberger 1998: 328–329, for example).

The underlying asset price is assumed to follow a conventional binomial process, as introduced by Cox *et al.* (1979). The asset price at time $t = 0$ is S, and rises exogenously, dependent perhaps but not necessarily on an asset pricing model such as the conventional CAPM (Nau and McCardle, 1991: 217). In the interval between $t = 0$ and $t = 1$, the price shifts either to S_u (it goes *Up*) or S_d (it goes *Down*).

The market price of the underlying asset at $t = 0$ is

$$S = \frac{\pi_u S_u + (1 - \pi_u) S_d}{(1 + r_s)}, \tag{12.1}$$

where π_u is the market's aggregated assessment of the probability of the asset going *Up* and r_s is the market (risk-adjusted) return on that asset.

Similarly, the market price at $t = 0$ of a call option with value at $t = 1$ equal to $C_u = Max(S_u - k, 0)$ given *Up* and $C_d = Max(S_d - k, 0)$ given *Down*, is

$$C = \frac{\pi_u C_u + (1 - \pi_u) C_d}{(1 + r_c)}, \tag{12.2}$$

where r_c is the market (risk-adjusted) rate of return on the call and k (a constant) is the strike or exercise price.

Up to this point, the derivation follows the 'naive' logic of expected present value, where the probabilities are 'actual' rather than 'risk neutral'. However, to find C knowing S, two further equations are necessary. These contain in static form the direct relationship between asset and derivative prices that the inventors of the risk-neutral option price, Black, Scholes and Merton, brought to financial decision theory.

The first comes from the observation that a portfolio containing

$$\Delta = \frac{C_u - C_d}{S_u - S_d} \tag{12.3}$$

shares and one call option sold short is risk free since its net value at $t = 1$ is the same whether the underlying asset value is S_u or S_d (*Up* or *Down*).[1] The second rests on a fundamental economic axiom, known as the 'law of no arbitrage' or the 'law of one price'. This says that any two assets which produce the same future cash flows under the same states of nature must have the same price, and therefore produce the same rate of return. The perfectly hedged portfolio defined by equation (12.3) replicates a risk-free bond and must therefore be priced such that it earns the risk-free rate of interest. That is, its weighted average return must equal r. Hence, the missing equation

$$\left(\frac{\Delta S}{\Delta S - C}\right) r_s + \left(\frac{-C}{\Delta S - C}\right) r_c = r. \tag{12.4}$$

Solving equations (12.1)–(12.4) simultaneously gives the now well-known formula for the call price

$$C = \frac{p_u C_u + (1 - p_u) C_d}{(1 + r)}, \tag{12.5}$$

[1]Equation (12.3) is found by setting the aggregate portfolio value under *Up*, $\Delta S_u - C_u$, equal to its value under *Down*, $\Delta S_d - C_d$, and solving for Δ.

where

$$p_u = \frac{S(1+r) - S_d}{S_u - S_d},$$

and $(1-p_u)$ are simplifying terms known as 'risk-neutral' probabilities (for the reason that if they were actual probabilities only someone who is risk neutral would value the option at C). The price of the call option is therefore its 'risk-neutral expectation' discounted at the risk-free rate. This same value C can be found from equation (12.2) using actual probabilities π and $(1-\pi)$, if indeed these are known, but only once the appropriate discount rate r_c is determined. Equation (12.4) provides the theoretical connection between r_c, r_s and r, however only r is observable if π is unknown.

The remarkable aspect of the solution (12.5) is that π does not appear, meaning that C can be found even when π (and hence r_s) is unknown, or the subject of disagreement between different individuals. Contrary to a common misunderstanding, this does not mean that π has no influence. To the contrary, π is accounted for by the market in the underlying asset price S. Different traders have different subjective estimates of π and these are aggregated in S. The call option price C is conditional on S and will change when a shift in the market's aggregate assessment of the actual probability π causes a reassessment of S. It is correct to say only that C is conditionally independent of π, or independent of π given S.

There are several misconceptions that have been promulgated widely in finance regarding risk-neutral valuation. The most fundamental of these is that in the context of standard decision trees, risk-neutral valuation methods are correct and neoclassical decision theory is wrong.

A full understanding of the relationship between these two decision frameworks is provided in Nau and McCardle (1991) and Smith and Nau (1995). In essence, risk-neutral valuation provides a direct and technically convenient (to say the least) method of valuing options that produces the same value of C (given S) as an individual expected utility maximizer would find by looking for Dutch Book investment combinations. If such combinations are not to exist, or not be handed by the investor to others, C must relate to S as per equation (12.5). Otherwise the decision-maker can reconstitute his investment portfolio so as to achieve a guaranteed increase (under either *Up* or *Down*) in *ex post* utility by selling an infinitesimal dollar amount of risk-free bonds and using the proceeds to finance a risk-free portfolio of shares and options (or *vice versa*).

Note that for such infinitesimal changes in portfolio weights, all investors, whatever their utility functions or risk aversion, are effectively risk-neutral. It is natural, therefore, that at the margin all investors see risk-neutral asset prices as 'correct' – that is, as mutually consistent, 'coherent' or arbitrage-free.

The principle of 'no Dutch book', or 'no arbitrage' as it is better known in finance, has existed explicitly in decision theory since de Finetti (1937). Its many important incarnations in modern finance (Nau and McCardle, 1991: 200) are testimony to de Finetti's insight. From this historical perspective, the solution to the option pricing puzzle was implicit and waiting to be discovered in classical expected utility maximization from the moment it was noticed that options could be combined in fixed proportions with stocks so as to achieve the very same risk-free outcome as (risk-free) bonds. That finance theorists reached the risk-neutral solution without explicit reference to expected utility is

indicative of the elegance of their economic logic rather than of any inherent defect in neoclassical economic decision theory.

12.3 Replicating investments with bets

A conventional bet of amount B on the event $\omega \in \Omega$, against the complementary event $\omega \notin \Omega$, is defined as as an outlay or 'investment' that returns a gross payoff equal to B multiplied by a factor

$$\begin{cases} \alpha_\Omega > 1 & \text{if } \omega \in \Omega \\ 0 & \text{if } \omega \notin \Omega. \end{cases}$$

For example, a bookmaker may quote a 'price' of $\alpha = 1.91$ on Connors to beat Borg. A gambler who bets x on Connors will be returned $1.91x$ if indeed Connors wins and zero if Borg wins. The win multiple α is always greater than one because it includes the dollar wagered.

In traditional British bookmaking language, the gross payoff (per \$1 bet) α_Ω from a successful bet on $\omega \in \Omega$ equals

$$\left[1 + \frac{1}{o_\Omega}\right] = \frac{1}{p_\Omega},$$

where $o_\Omega = \frac{p_\Omega}{1-p\Omega}$ represents the odds 'in favour' of event $\omega \in \Omega$ (often called 'odds on') and p_Ω is the odds-implied probability of $\omega \in \Omega$. In the absence of any bookmaker commission or spread, the reciprocal of o_Ω represents the odds in favor of $\omega \notin \Omega$ (odds against $\omega \in \Omega$) and the odds-implied probability of $\omega \notin \Omega$ is $1 - p_\Omega$.

Note that there is no spread in a conventional binomial tree. The assumption is that an investor can either buy or sell the underlying asset (or any of its derivatives) at a given price.

Investments can be replicated with bets as follows. Consider an investment position long one unit in the underlying asset. This is replicated by a portfolio containing $\frac{S_d}{1+r}$ in risk-free bonds together with a bet of amount $S - \frac{S_d}{1+r} = \frac{S(1+r)-S_d}{1+r}$ on *Up* at payoff

$$\alpha_u = \left[1 + \frac{1}{o_u}\right](1+r) = \frac{(1+r)}{p_u} = \frac{S_u - S_d}{S(1+r) - S_d}(1+r), \tag{12.6}$$

where $o_u = \frac{p_u}{1-p_u}$ represents the risk-neutral odds in favour of *Up*. The gambler's net outlay at $t = 0$ is then S, and the total value of his portfolio of bonds-plus-bet at $t = 1$ is S_u in the case of *Up* and S_d in the case of *Down*, the same as if he had bought one unit of the underlying asset.

Now consider a position short one unit in the underlying asset. This is replicated by a portfolio containing $-\frac{S_u}{1+r}$ (a short position) in risk-free bonds together with a bet of amount $\frac{S_u}{1+r} - S = \frac{S_u - S(1+r)}{1+r}$ on *Down* at payoff

$$\alpha_d = \left[1 + \frac{1}{o_d}\right](1+r) = \frac{(1+r)}{1-p_u} = \frac{S_u - S_d}{S_u - S(1+r)}(1+r), \tag{12.7}$$

where $o_d = \frac{1}{o_u} = \frac{1-p_u}{p_u}$ represents the risk-neutral odds in favour of *Down*.

The gambler's net outlay at $t = 0$ is then $-S$, meaning that he receives amount S, the same as if he had sold one unit of underlying asset, and the total value of his portfolio of bonds-plus-bet at $t = 1$ is $-S_u$ in the case of *Up* and $-S_d$ in the case of *Down*. This exactly replicates the short sale of one unit of the underlying asset.

12.4 Log optimal (Kelly) betting

Consider an investor whose decision rule is to maximize the expected log utility of wealth at some future time T, $E[Log(wealth_T)]$. This is a myopic decision rule since $E[Log(wealth_T)]$ is maximized by maximizing $E[Log(wealth_t)]$ at each time t prior to T (Luenberger 1998: 426). A bet or set of bets designed to maximize expected log utility is known to professional gamblers, following Kelly (1956), as 'Kelly betting', and is much celebrated (Poundstone, 2005).

In finance, expected log utility maximization is known as 'log optimal' investment and was introduced by Hakansson (1971), Roll (1973), Markowitz (1976) and Rubinstein (1976). First among its many interesting properties (*cf.* Maclean *et al.* 1992; Luenberger 1998), Kelly (1956) showed that maximizing $\bar{U} = E[Log(wealth_t)]$ implicitly maximizes the decision-maker's exponential capital growth over the interval preceding t, as $t \to \infty$. Gamblers who believe they have an edge (e.g. card counters in Blackjack) bet this way so as to obtain the greatest possible long run average increase in their bankroll, and thus exploit their edge to maximum monetary effect (Thorp 1966, 1969; Luenberger 1998: 429). In economics, Blume and Easley (1992, 2001, 2002) demonstrated that financial markets, reduced to their basic elements, 'select for' those rational decision-makers (expected utility maximizers) with log utility and physically 'true' probabilities. In the long run, a decision-maker with these twin attributes will ultimately ruin any other, and one with log utility will ruin another with less accurate probabilities regardless of the other's utility function.

12.4.1 Kelly betting in a binomial tree

Suppose that a Kelly (log utility) gambler with starting wealth W believes that the probability of *Up* is q_u. Following Kelly (1956: 922), the gambler bets a fixed proportion γ of his wealth on *Up* and the remainder $(1-\gamma)$ on *Down*. Because there is no commission (spread), these bets are partly self-cancelling, meaning that there is implicitly a proportion of the gambler's initial wealth that remains unbet and invested at the risk-free rate r. The gambler's expected utility is

$$q_u\ Log\left[\gamma W\frac{(1+r)}{p_u}\right]+(1-q_u)\ Log\left[(1-\gamma)\,W\frac{(1+r)}{(1-p_u)}\right]$$

$$= Log[W(1+r)]+q_u\ Log\left(\frac{\gamma}{p_u}\right)+(1-q_u)\ Log\left(\frac{1-\gamma}{1-p_u}\right).$$

Differentiating with respect to γ leads to a maximum such that

$$\frac{q_u}{\gamma}-\frac{(1-q_u)}{1-\gamma}=0,$$

giving $\gamma = q_u$ and $(1-\gamma) = (1-q_u)$. It follows therefore that a Kelly gambler or log optimal investor must allocate his initial wealth, whatever its total amount, to matching bets on *Up* and *Down* in exact proportion to his subjective probabilities of those events, q_u and $(1-q_u)$, regardless of the respective payoffs α_u and α_d. The Kelly bettor's wealth in the event of *Up* is then $W(1+r)\frac{q_u}{p_u}$, and in the event of *Down*, $W(1+r)\frac{1-q_u}{1-p_u}$.

Thus, an appealingly simple result occurs. Specifically, the gambler's wealth after Kelly betting equals his prior wealth multiplied by a factor

$$\frac{q}{p}(1+r), \tag{12.8}$$

where q represents his (*ex ante*) subjective probability of the event actually observed (*Up* or *Down*) and p represents the risk-neutral probability of that event implied by the market.

Gamblers might be offered this payoff function as a universal 'Kelly derivative' of the underlying asset. To invest in a Kelly derivative, the gambler is required to stipulate two quantities at $t = 0$, namely (1) his personal or nominated probability q_u of *Up* (implying $q_d = 1 - q_u$) and (2) his wealth, or the part thereof that he wants to invest growth-optimally.

To profit from such an investment, relative to merely holding all wealth at $t = 0$ in risk-free bonds, the gambler must predict the eventual outcome *Up* or *Down* with higher probability than 'the market'. For example, if he places personal probability 0.9 on that event when the market offers a risk-neutral probability of 0.6, then his total wealth at $t = 1$ will be 50% higher than if he had meekly invested everything in riskless bonds.

12.4.2 Kelly derivatives

Suppose that we define a '$1 Kelly bet' as a derivative of the underlying asset that costs $1 and pays

$$\frac{q}{p}(1+r) = \begin{cases} \frac{q_u}{p_u}(1+r) & \text{if } UP \\ \frac{1-q_u}{1-p_u}(1+r) & \text{if } Down, \end{cases}$$

where (as above) q_u and p_u are the gambler's and market's quoted probabilities of *Up*, respectively. The term $(1+r)$ appears in these expressions whenever the 'market probability' is a risk-neutral probability. Risk-neutral probabilities are derived on the basis that all assets return the risk-free rate. By 'outpredicting the market' (i.e. by nominating $q > p$), the gambler gains an extra return over and above the risk-free rate.

The gambler can take up as many $1 Kelly bets or 'Kelly derivatives' as he desires. If he is a strict Kelly bettor, he will invest his entire wealth (i.e. if he has $100, he will buy 100 $1 bets). More typically, he may have a bankroll set aside for gambling, and treat the remainder of his fortune (e.g. his house) as a separate 'mental account' in the behavioural finance sense, in which case he will invest 100% of that smaller and somewhat arbitrary amount.

This does not mean that he can lose 100%. If he is a bad judge, or merely unlucky, q will be small relative to p and he will be left with just a small proportion $\frac{q}{p}(1+r)$ of each

\$1 bet, but provided he does not ever nominate a personal probability q_u equal to 0 or 1, this factor will always exceed zero.

A further possibility is that the gambler might invest in a weighted portfolio of '\$1 Kellys', some with different nominal values of q_u (p_u is fixed by the market). For example, he could buy \$20 worth of bets at $q_u = 0.9$ and \$80 worth of bets at $q_u = 0.6$, thereby 'hedging his bets'. The $q_u = 0.9$ bets would payoff extra well in the event of *Up* but lose more heavily in the event of *Down*.

Note that a true Kelly bettor would not do this. By hedging, or betting on any probability other than his actual belief, he would reduce his expected long-run capital growth and hence not meet his implicit objective. More specifically, his subjective expected utility from Kelly betting some proportion θ ($0 \leq \theta \leq 1$) of an initial endowment of \$1 *as if* his personal probability of *Up* was g_u, when actually it is q_u, is

$$q_u Log\left[\left\{\theta\frac{g_u}{p_u} + (1-\theta)\right\}(1+r)\right] + (1-q_u)\,Log\left[\left\{\theta\frac{1-g_u}{1-p_u} + (1-\theta)\right\}(1+r)\right]$$

which is maximized when $\theta = 1$ and $g_u = q_u$. It follows, therefore, that if the gambler truly has a log utility function, and provided that he buys Kelly bets at a nominal probability of say $q_u = 0.81$, then 0.81 is his true degree of belief. (The total number of \$1 bets purchased at this probability depends on his wealth.) 'Kelly profit' is thus a 'proper scoring rule' in the sense of de Finetti (1974). That is, the probability forecaster (here gambler) maximizes his subjective expected 'score' by nominating his true subjective degree of belief.

12.5 Replicating Kelly bets with puts and calls

A Kelly bet or log optimal investment in the underlying asset can be replicated with a portfolio of put options, call options and risk-free bonds. Let the value of a call at $t = 1$ be $C_u = Max[S_u - k_c,\ 0]$ when $S = S_u$, and $C_d = Max[S_d - k_c,\ 0]$ when $S = S_d$, where k_c is the strike price. Similarly, let the value of a put at $t = 1$ equal $P_u = Max[k_p - S_u, 0]$ when $S = S_u$, and $P_d = Max[k_p - S_d, 0]$ when $S = S_d$, where k_p is the strike price.

Now, consider a portfolio with weight θ_c in calls, θ_p in puts, and θ_{rf} in riskless bonds, with

$$\theta_c + \theta_p + \theta_{rf} = 1. \tag{12.9}$$

To replicate a Kelly bet, this portfolio must satisfy the joint conditions

$$\theta_c\frac{C_u}{C} + \theta_p\frac{P_u}{P} + \theta_{rf}(1+r) = \frac{q_u}{p_u}(1+r) \tag{12.10}$$

and

$$\theta_c\frac{C_d}{C} + \theta_p\frac{P_d}{P} + \theta_{rf}(1+r) = \frac{1-q_u}{1-p_u}(1+r). \tag{12.11}$$

Equation (12.9) is implicit within equations (12.10) and (12.11) and is therefore redundant, as can be seen by substituting for $C = \frac{p_u C_u + (1-p_u)C_d}{1+r}$ and $P = \frac{p_u P_u + (1-p_u)P_d}{1+r}$, and adding

equations (12.10) to (12.11). Solving equations (12.10) and (12.11) simultaneously and eliminating θ_{rf} leads to the simple condition

$$\theta_c = \frac{q_u - p_u(1-\theta_p)}{1-p_u}. \tag{12.12}$$

Any portfolio of puts, calls and riskless bonds satisfying this condition, together with $\theta_{rf} = 1-(\theta_c+\theta_p)$, replicates a log optimal investment (Kelly bet) in the underlying asset. It is possible, therefore, to set either θ_c or θ_p arbitrarily. Furthermore, either can be zero, meaning that a log optimal portfolio may contain only calls and bonds, or just puts and bonds. With $\theta_c = 0$, equation (12.12) becomes $\theta_p = \frac{p_u - q_u}{p_u}$, and with $\theta_p = 0$, $\theta_c = \frac{q_u - p_u}{1-p_u}$. The most remarkable aspect of these conditions, including equation (12.12), is that the log optimal portfolio weights are fully determined by the rival probabilities, p_u and q_u.

12.6 Options on Kelly bets

A Kelly bet is a binary asset like any other, in the sense that it has an uncertain value at $t = 1$ that may take one of two values. It is possible, therefore, to define and price an option on a \$1 Kelly bet. Let this be a call option with value at $t = 1$

$$\begin{cases} C_u = Max\left[\frac{q_u}{p_u}(q+r)-k, 1\right] & \text{if } Up \\ C_d = Max\left[\frac{1-q_u}{1-p_u}(1+r)-k, 1\right] & \text{if } Down, \end{cases}$$

where k is the strike price. The risk-neutral (no-arbitrage) value of the option is given by equation (12.5), where C_u and C_d are as defined above. Assuming, for example, that $q_u > p_u$, and $k \le r$, this option is in the money given Up (i.e. $C_u > 1$) and out of the money given $Down$, (i.e. $C_d = 1$). In this case,

$$C = \frac{1 + q_u(1+r) - p_u(1+k)}{1+r}. \tag{12.13}$$

A numerical example is as follows. Let the gambler's nominated probability of Up be $q_u = 0.9$ and the risk-neutral probability of Up be $p_u = 0.3$. Assuming for example that $r = 0.1$ and $k = 0.20$, the price at $t = 1$ of the call option is $C_u = \frac{0.9}{0.3}(1.10) - 0.2 = \$3.10$ given Up, or $C_d = \$1$ given $Down$. Substituting in equations (12.5) or (12.13), its value at $t = 0$ is \$1.48. Thus, it costs \$1.48 at time $t = 0$ to buy a call option on a \$1 Kelly bet on Up with nominal $q_u = 0.9$ and exercise price $k = 0.20$.

12.6.1 Log optimal investment in options on Kelly bets

Log optimal investment in options on \$1 Kelly bets is no different in principle to log optimal investment in any other binary asset. Taking a long position in a call option on

a \$1 Kelly bet amounts to betting some chosen sum on *Up* at gross payoff (per dollar wagered)

$$\beta_u = \frac{C_u - C_d}{C(1+r) - C_d}(1+r) = \frac{(1+r)}{p_u}.$$

Similarly, taking a short position in a call option amounts to betting on Down at gross payoff

$$\beta_d = \frac{C_u - C_d}{C_u - C(1+r)}(1+r) = \frac{(1+r)}{1-p_u}.$$

These are the same payoffs as equations (12.6) and (12.7). Again there is no spread between the prices at which a call option is bought or sold, and hence the standard Kelly result applies. That is, a log optimal investment portfolio is achieved by making matching bets on *Up* and *Down* in proportion to the gambler's subjective probabilities of these events (regardless of the respective payouts of the two events).

There is no difference, therefore, between log optimal investment in the underlying asset (here \$1 Kelly bets) and in options on that asset. In either case, the log optimal investor's wealth at $t = 1$ is given by equation (12.8) and depends only on the accuracy of his confessed probability q of the observed event (*Up* or *Down*) relative to the market-implied risk neutral probability p of that outcome. This is true of log optimal investment in any conceivable derivative of the same underlying asset.

12.7 Conclusion

Betting or prediction markets and financial markets have now largely converged in their microstructure. The most telling shift within betting markets has been the advent of internet-based betting exchanges, on which gamblers can both buy ('back') and sell ('lay') any chosen security (e.g. horse), provided there are willing counterparties (Smith *et al.*, 2006: 674). The order book on a betting exchange such as *Betfair* is essentially identical to that on any order-driven stock exchange, such as SETS on the London Stock Exchange. A recent paper by Ozgit (2005) describes the convergence of betting and gambling markets, and contains a 'market microstructure' analysis of *Betfair* betting exchange data founded on previous studies in financial market microstructure.

It is likely that as the parallels between physical and on-line markets for bets and those for more conventional financial assets such as stocks and options become widely understood, and as the day-to-day financial turnover on international betting or 'prediction' markets rivals that on long-life asset markets, there will be a merger of the two related mathematical and empirical research literatures. This chapter is intended to contribute to such 'cross-disciplinary' consensus by clarifying the implicit equivalence of trading stocks (or derivatives) within binomial asset pricing models and simply betting on *Up* or *Down*.

The most interesting aspect of this analogy is the part played by 'risk-neutral' rather than 'actual' probabilities. Unlike natural 'physical' or 'personal' probabilities, risk-neutral probabilities do not occur in conventional gambling mathematics. A further and apparently original suggestion is that a 'Kelly bet' on the underlying asset is replicated very

simply by a portfolio of riskless bonds and derivatives of that asset (either, or both, calls or puts).

By better understanding the betting-investment analogy, both the financial markets and betting industries might learn from one another. It is feasible that the betting market for 'exotics' (e.g. parlays and point spread bets) will mimic the growth and complexity of financial derivatives markets. Similarly, financial market makers have already begun to offer products such as 'binaries' (e.g. bets on whether the stock market goes up on the day) that may appeal as much to the traditional High Street betting community as to their 'more sophisticated' clientelle. There is much scope for business development in both markets, and also the possibility of firms re-engineering their products either as 'bets' or 'investments', depending on which attracts the more favourable regulatory and taxation treatments in the countries concerned.

Ultimately, particularly if some betting markets offer systematic inefficiencies, it is conceivable that financial institutions might establish betting market hedge funds. In principle, apart from the historical disparities in transactions costs and liquidity, there is no difference between betting on stocks and football matches. Both markets contain more or less informed traders, each trying to out-predict one another, and a cohort of noise or liquidity traders, some of whom believe they are informed when they are not, and some of whom are just having a flutter for the thrill of it or for almost any psychological or sociological reason imaginable (e.g. to impress or mimic their peers). Moreover, sports and other betting events, such as the winners of the academy awards, are statistically appealing in that they offer returns with presumably zero covariance with other assets, and hence zero 'beta'.[2] According to the capital asset pricing model, their required return, if they are to warrant some small place in a well diversified investment portfolio, is just the risk-free rate.

References

Asch, P., Malkeil, B. G. and Quandt, R. E. (1982). Racetrack betting and informed behavior. *Journal of Financial Economics*, 10:187–194.

Blume, L. and Easley, D. (1992). Evolution and market behavior. *Journal of Economic Theory*, 58:9–40.

Blume, L. and Easley, D. (2001). If you're so smart, why aren't you rich? Belief selection in complete and incomplete markets. Cowles Foundation Discussion Paper No. 1319. Yale University.

Blume, L. and Easley, D. (2002). Optimality and natural selection in markets. *Journal of Economic Theory*, 107:95–135.

Boyle, P. P. (1992). *Options and the Management of Financial Risk*. Schaumburg, IL.: Society of Actuaries.

Cox, J. C. and Rubinstein, M. (1985). *Options Markets*. Englewood Cliffs. NJ: Prentice Hall.

Cox, J. C., Ross, S. A. and Rubinstein, M. (1979). Option pricing: a simplified approach. *Journal of Financial Economics*, 7:229–263.

de Finetti, B. (1937). La Prévision; ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1–68. Reprinted as 'Foresight: Its logical laws, its subjective sources', in H. E. Kyburg and H. E. Smokler (eds), *Studies in Subjective Probability*, 2nd edn (1980). New York: Kreiger, pp. 54–118.

de Finetti, B. (1974). *Theory of Probability*, Vol. 1. New York, NY: Wiley.

Edmans, A., Garcia, D. and Norli, O. (2006). Sports sentiment and stock returns. *Journal of Finance*, in press.

[2]This is naturally true of horse races and sports in general, but not perhaps of all sports events. Recent evidence (Edmans *et al.*, 2006) suggests that emotional reaction to the outcomes of very major international sports events may be reflected in the stock markets of participating countries, with wins promoting positive sentiment and price increases.

Gandar, J. M., Dare, W. H., Brown, C. R. and Zuber, R. A. (1998). Informed traders and price variations in the market for professional basketball games. *Journal of Finance*, 53:385–401.
Golec, J. and Tamarkin, M. (1991). The degree of inefficiency in the football betting market: statistical tests. *Journal of Financial Economics*, 30:311–323.
Gray, P. K. and Gray, S. F. (1997). Testing market efficiency: evidence from the sports betting market. *Journal of Finance*, 52:1725–1737.
Hakansson, N. H. (1971) Capital growth and the mean-variance approach to portfolio selection. *Journal of Financial and Quantitative Analysis*, 6:517–557.
Hausch, D. G. and Ziemba, W. T. (1990). Arbitrage strategies for cross-track betting on major horseraces. *Journal of Business*, 63:61–78.
Hausch, D. G., Ziemba, W. T. and Rubinstein, M. E. (1981). Efficiency of the market for racetrack betting. *Management Science*, 27:1435–1452.
Hausch, D. G., Lo, V. and Ziemba, W. T. (eds) (1994). *Efficiency of Racetrack Betting Markets*. New York, NY: Academic Press.
Ingersoll, J. E. (1987). *Theory of Financial Decision Making*. Savage, MD: Rowman and Littlefield.
Kelly, J. (1956). A new interpretation of the information rate. *Bell System Technical Journal*, 35:917–926.
Levitt, S. D. (2004). Why are betting markets organized so differently from financial markets? *The Economic Journal*, 114:223–246.
Luenberger, D. (1998). *Investment Science*. New York, NY: Oxford University Press.
MacLean, L. C, Ziemba, W. T. and Blazenko, G. (1992). Growth versus security in dynamic investment analysis. *Management Science*, 38:1562–1585.
Markowitz, H. M. (1976) Investment for the long run: new evidence for an old rule. *Journal of Finance*, 31:1273–1286.
Nau, R. F. and McCardle, R. F. (1991). Arbitrage, rationality and equilibrium. *Theory and Decision*, 31:199–240.
Ozgit, A. (2005). The bookie puzzle: auction versus dealer markets in online sports betting. Working Paper, Department of Economics, UCLA.
Pope, P. F. and Peel, D. A. (1989). Information, prices and efficiency in a fixed-odds betting market. *Economica*, 56:323–341.
Poundstone, W. (2005). *Fortune's Formula: The Untold Story of the Scientific Betting System that Beat the Casinos and Wall Street*. New York, NY: Farrar, Straus and Giroux.
Roll, R. (1973). Evidence on the growth optimum model. *Journal of Finance*, 28:551–567.
Rubinstein, M. (1976). The strong case for the generalized logarithmic utility model as the premier model of financial markets. *Journal of Finance*, 31:551–571.
Sauer, R. D. (1998). The economics of wagering markets. *Journal of economic Literature*, 36:2021–2064.
Shin, H. S. (1993). Measuring the incidence of informed trading in a market for state-contingent claims. *The Economic Journal*, 103:1141–1153.
Smith, T. E. and Nau, R. F. (1995). Valuing risky projects: option pricing theory and decision analysis. *Management Science*, 41: 79–816.
Smith, M. A., Paton, D. and Vaughan Williams, L. (2006). Market efficiency in person-to-person betting. *Economice*, 73:673–689.
Stoll, H. R. and Whaley, R. E. (1993). *Futures and Options: Theory and Applications*. Cincinatti, OH: South Western.
Thaler, R. and Ziemba, W. (1988). Parimutual betting markets: racetracks and lotteries. *Journal of Economic Perspectives*, 2:161–174.
Thorp, E. (1966). *Beat the Dealer*, 2nd edn. New York, NY: Vintage.
Thorp, E. (1969). Optimal gambling systems for favorable games. *International Statistical Review*, 37:273–293.
Thorp, E. (1971). Portfolio choice and the Kelly criterion. *Proceedings of the Business and Economics Section of the American Statistical Association*, pp. 215–224. Reprinted in Ziemba, W. T. and Vickson, R. G. (1975). *Stochastic Optimization Models in Finance*. New York, NY: Academic Press, pp. 599–619.
Thorp, E. (2000). The Kelly criterion in blackjack, sports betting and the stock market. In: O. Vancura, J. Cornelius, and W. R. Eadington (eds), *Finding the Edge: Mathematical Analysis of Casino Games*. Reno, NV: Institute for the Study of Gambling and Commercial Gaming, pp. 163–213.
Vaughan Williams, L. (1999). Information efficiency and betting markets. *Bulletin of Economic Research*, 51:1–30.
Vaughan Williams, L. (2005). *Information Efficiency in Financial and Betting Markets*. Cambridge: Cambridge University Press.

Vaughan Williams, L. and Paton, D. (1997). Why is there a favourite-longshot bias in British racetrack betting markets? *The Economic Journal*, 107:150–158.
Woodland, B. and Woodland, L. (1994). Market efficiency and the favorite-longshot bias: the baseball betting market. *Journal of Finance*, 49:269–279.
Ziemba, W. T. and Hausch, D. (1985). *Betting at the Racetrack*. Los Angeles, CA: Dr Z Investments.

13 The hidden binomial economy and the role of forecasts in determining prices

Stephen Satchell and Oliver Williams

Abstract

The purpose of this chapter is to investigate the many possible versions of underlying economies that are consistent with the usual binomial structure central to option pricing. We use these economies to model explicitly a number of interesting problems that are usually intractable in more complex frameworks. In particular, we look at equilibria where agents have subjective expected utility and heterogeneous utility functions that do not satisfy standard aggregation conditions. We show that in such a one-period world, price is driven entirely by forecasts and has no connection with true probability. We derive formulae for equilibrium equity prices in a number of novel cases.

13.1 Introduction

Very little is known about the impact on market prices of different beliefs/forecasts. Intuitively, it is thought that heterogeneity in forecasts should be a good thing in that it reduces market volatility, ensuring active trading among agents and hence supporting liquidity. However, there is a scarcity of research about economies where agents differ in both their beliefs and utility. Most closed-form expressions for single and aggregate demands involve fairly strong assumptions about homogeneity. There are some studies in this area – see, for example, Constantinides (1982) and Hara *et al.* (2006a, 2006b). These authors consider fairly generic properties of heterogeneous equilibria; they do not focus on explicit forms for individual demands. Numerous authors (see Chiarella *et al.*, 2006) consider heterogeneous Capital Asset Pricing Models, but there the difference is in beliefs and risk aversion rather than, for example, different indexed HARA[1] utilities – the point being that the degree of heterogeneity is not especially large.

In particular, analysis of the most empirically plausible case of an economy in which investors each have their own beliefs and unrestricted HARA utilities presents challenges both technical and semantic. On the technical front, assuming agents are von Neumann–Morgenstern expected-utility maximizers, the support of each agent's returns distribution must be compatible with the form of the particular agent's utility function (e.g. a power utility agent cannot have normally-distributed beliefs over future asset prices, since this would imply a strictly positive probability of negative future wealth which would

[1] A utility function $U(w)$ belongs to the Hyperbolic Absolute Risk Aversion (HARA) class if $-\frac{U''(w)}{U'(w)} = \frac{1}{A+Bw}$

have undefined or imaginary utility). Additionally, even if analytic expressions for each expected utility are available, solving the resulting equilibrium conditions may require numerical methods. From a semantic point of view, while the notion of a representative agent is quite intuitive in the case of homogeneous beliefs, introducing heterogeneous beliefs complicates matters somewhat. As highlighted by Rubinstein (1974), in the limited set of HARA cases where aggregation results *are* available, the representative agent must also have suitably-defined 'composite' beliefs. Although this does not detract from the power of the result, it may make it less transparent to draw conclusions about aggregate risk aversion from the properties of representative utility alone.

None of these would seem important in the wider world, except that central banks and regulators routinely use a representative agent as a barometer of market risk.

Our chapter looks at an economy that underpins the binomial option pricing model, generalizing somewhat the setting of Johnstone (2006). This model is interesting because the focus is always on the price of the redundant asset (the derivative) and not the price of the traded asset (the stock), which is always taken as given. However, it is worth studying, not least because its simple structure allows one to look at problems that are usually intractable in more complex models. For instance, since the traded asset return is by definition a Bernoulli event, an agent's *beliefs* can be completely characterized by a single probability number rather than a continuous density. This makes for easy interpretation of heterogeneity as well as a straightforward expected utility computation.

In our models, agents are allowed to have various combinations of different beliefs and different indexed HARA and other utilities. These beliefs take the form of forecasts. We derive a number of corresponding equilibrium asset prices, in closed form. The chapter is organized as follows: Section 13.2 outlines the general structure of the models, and Section 13.3 considers cases of power utility where we get an explicit expression for the equilibrium price in terms of wealth, possible Bernoulli outcomes, subjective probabilities, coefficients of risk aversion and the riskless rate. It is notable that the true probabilities do not enter into the equilibrium. In Section 13.4 we consider cases of what we term *mixed equilibria*; these are situations where individuals are fully heterogeneous in that utility functions come from different families – for example, we analyze the interesting problem of a rational power utility investor trading with an individual who practises prospect theory. Section 13.5 concludes.

13.2 General set-up

We consider an economy consisting of a riskless asset (bond) and a single risky asset. Agents allocate their wealth at time-0 between these two assets according to their subjective beliefs. Wealth, in the context of this model, might be thought of as units of time-1 consumption from which no utility can be derived at time-0. However, it is straightforward to add a time-0 utility to the problem, and we shall do so at various points. At time-1 there are only two possible states of the world: an 'up' state, in which the risky asset has a rate of return u, and a 'down' state, in which it has a rate of return d. The rate of return on the risk-free asset is r_f. We denote by f_i the subjective probability (in the opinion of agent i) that the time-1 state will be 'up'. As usual, we require that $u > r_f > d$.

The risk-neutral probability of the 'up' state is denoted π and therefore defined as follows:

$$\pi = \frac{r_f - d}{u - d} \tag{13.1}$$

with

$$\frac{u - r_f}{r_f - d} = \frac{1 - \pi}{\pi}$$

We also define the odds ratio

$$h_i = \frac{\dfrac{f_i}{1 - f_i}}{\dfrac{\pi}{1 - \pi}} \geq 0 \tag{13.2}$$

which indicates the relative optimism of agent i with respect to the risk-neutral probability. In fact, we shall call this parameter *optimism* in what follows.

At time-0, the wealth which agent i has available for investment is denoted w_{0i} and aggregate investible wealth is denoted $W_0 \equiv \sum_i w_{0i}$. There are $\overline{m}$ units of the riskless bond outstanding in the economy (each of which redeems at time-1 at a value of 1 wealth unit) together with $\overline{n}$ units of the risky asset. The share of wealth which agent i allocates to the risky asset will be denoted x_i, the equilibrium spot risky asset price S_0 and equilibrium bond price B_0.

The following relationship must therefore apply in equilibrium, and can be used to solve for r_f and hence S_0:

In the case where $\overline{m} = 0$ (no bonds outstanding), then of course

$$\sum_{i=1}^{H} w_{0i} = \sum_{i=1}^{H} w_{0i}\, x_i = \overline{n} S_0$$

$$S_0 = \frac{1}{\overline{n}} \sum_{i=1}^{H} w_{0i}$$

and the bond is not identified. Otherwise we have

$$B_0 = \frac{1}{1 + r_f} \tag{13.3}$$

$$\begin{aligned} \frac{1}{1 + r_f} &= \frac{1}{\overline{m}} \sum_{i=1}^{H} w_{0i} - \frac{1}{\overline{m}} \sum_{i=1}^{H} w_{0i}\, x_i \\ &= \frac{1}{\overline{m}} \sum_{i=1}^{H} w_{0i} - \frac{\overline{n}}{\overline{m}} S_0 \end{aligned} \tag{13.4}$$

$$r_f = \frac{1}{\sum_{i=1}^{H} \frac{w_{0i}}{\overline{m}} - \frac{\overline{n}}{\overline{m}} S_0} - 1$$

In general, risky asset demand x_i will itself depend on r_f, and hence determining this riskless interest rate is a necessary first step in solving for a full equilibrium. Given r_f, however, it is straightforward to determine S_0. In this chapter we consider the conditional problem of determining equity price *given* interest rates, and we solve the full problem in a separate paper.

13.2.1 State prices

Our economy consists of two assets and has two states. The payout matrix is:

	Asset 1	**Asset 2**
State 1	$S_0(1+u)$	1
State 2	$S_0(1+d)$	1

So in order to compute the state 1 price, for instance, we require to find weightings on each asset (stock and bond) such that

$$\lambda_1(\text{asset1}) + \lambda_2(\text{asset2})$$
$$= 1 \text{ in state 1}$$
$$= 0 \text{ in state 2}$$

hence

$$\lambda_1 S_0(1+u) + \lambda_2 1 = 1$$
$$\lambda_1 S_0(1+d) + \lambda_2 1 = 0$$
$$\Rightarrow \quad \lambda_1 S_0(u-d) = 1$$
$$\lambda_1 = \frac{1}{S_0(u-d)}$$
$$\lambda_2 = 1 - \frac{S_0(1+u)}{S_0(u-d)}$$
$$= \frac{-S_0(1+d)}{S_0(u-d)}$$
$$= -\frac{1+d}{u-d}$$

Therefore the state 1 price is simply

$$\begin{aligned} &= \frac{1}{S_0(u-d)} S_0 - \frac{(1+d)}{u-d} \frac{1}{1+r_f} \\ &= \frac{r_f - d}{(u-d)(1+r_f)} \\ &= \frac{\pi}{1+r_f} \end{aligned} \tag{13.5}$$

i.e. the risk-neutral probability discounted at the riskless rate. This is entirely as expected in a complete market economy such as this. Since a portfolio consisting of both state securities will return 1 in either 'up' or 'down' state, it is clear that by arbitrage such a portfolio must cost the same as the riskless bond, i.e. $B_0 = \frac{1}{1+r_f}$. Hence state 2 price is trivially given by $\frac{1-\pi}{1+r_f}$.

13.2.2 Asset demands

We shall now go a stage further and characterize investors each in terms of their time-0 holdings of assets. We have denoted the wealth of agent i by w_{0i}. This in turn consists of an initial endowment of $\overline{n}_i$ shares and $\overline{m}_i$ bonds, and hence

$$w_{0i} = \overline{n}_i S_0 + \overline{m}_i/(1+r_f) \tag{13.6}$$

We assume that agent i has time-0 utility $u_i(c_{0i})$ and time-1 utility $u_i(w_{1i})$. Agent i invests $I_i = w_{0i} - c_{0i}$, in n_i shares and m_i bonds. It follows that

$$I_i = n_i S_0 + m_i/(1+r_f)$$

and

$$w_{1i} = n_i S_1 + m_i \tag{13.7}$$

If we eliminate our budget constraint, then agent i chooses c_{0i} and n_i to maximize the following.

$$\begin{aligned} V_i &= u_i(c_{0i}) + E_{0i}(u_i(w_{1i})) \\ &= u_i(c_{0i}) + E_{0i}(u_i(n_i S_1 + (1+r_f)(I_i - n_i S_0))) \\ &= u_i(c_{0i}) + E_{0i}(u_i(n_i(S_1 - S_0(1+r_f)) + (1+r_f)I_i)) \end{aligned}$$

First-order conditions are therefore

$$\frac{\partial V_i}{\partial c_{0i}} = u_i'(c_{0i}) - E_{0i}(u_i'(w_{1i}))(1+r_f) = 0$$

and

$$\frac{\partial V_i}{\partial n_i} = E_{0i}(u_i'(w_{1i})(S_1 - S_0(1+r_f))) = 0 \tag{13.8}$$

We note that $E_{0i}(.)$ means taking expectations with respect to f_i and $1-f_i$.

Equation (13.8) implies that:

$$u_i'(n_iS_0(1+u)+m_i)(u-r_f)f_i = u_i'(n_iS_0(1+d)+m_i)(r_f-d)(1-f_i)$$

or, in terms of agent's *optimism*:

$$\frac{u_i'(n_iS_0(1+d)+m_i)}{u_i'(n_iS_0(1+u)+m_i)} = h_i \tag{13.9}$$

following from our definition of h_i in equation (13.2).

In passing, it is worth noting that the optimality conditions above can equivalently be expressed state-by-state as:

$$u_i'(n_iS_0(1+u)+m_i)f_i = \lambda\pi$$

$$u_i'(n_iS_0(1+d)+m_i)(1-f_i) = \lambda(1-\pi)$$

where

$$\lambda = \frac{u_i'(c_{0i})}{1+r_f} = E_{0i}(u_i'(w_{1i}))$$

is a Lagrangean multiplier which can be interpreted as an expected marginal utility of wealth. This is the framework used by Varian (1985), among others, which emphasizes the relationship between state prices and beliefs f_i and can be generalized to many (or continuous) states. A more detailed derivation of this equivalence appears in the Appendix to this chapter.

Equation (13.9) has the following implications:

Proposition 1: If $u_i(\,)$ is increasing and concave, ($u_i'(\,) > 0$ and $u_i''(\,) < 0$) then

$h_i > 1$ iff $n_i > 0$

$h_i = 1$ iff $n_i = 0$

$0 < h_i < 1$ iff $n_i < 0$.

Proof: This follows immediately from noting that $n_i > 0$ iff $n_iS_0(1+u)+m_i > n_iS_0(1+d)+m_i$ iff $u'(n_iS_0(1+u)+m_i) < u'(n_iS_0(1+d)+m_i)$; other arguments take the same form.

Remark 1: Proposition 1 has the corollary that any agent who is 'risk-neutral in beliefs' will set $n_i = 0$, and thus will not hold equity and sell his endowment $\overline{n}_i$ (assumed non-negative). However, individuals who are risk neutral in utility can only be in equilibrium in the sense of equation (13.8) if they have risk-neutral beliefs, otherwise they will wish to go infinitely long or short. Given a computed n_i for each i, equilibrium requires that $\sum_{i=1}^{H} n_i = \sum_{i=1}^{H} \overline{n}_i$. However to proceed further, we need to consider specific choices for $u_i(\,)$. If our only interest is the determination of the equity price, then we need only use equation (13.9).

Remark 2: It also follows that we can express 'trade' for equities and bonds by agent i as $n_i - \overline{n}_i$ and $m_i - \overline{m}_i$.

13.3 Power utility

In this section we consider our earlier analysis in the special case where our H agents all have different power utility functions; in particular, they differ in investible wealth, risk aversion and optimism.

Agent i acts to maximize expected utility given by:

$$V_i = \frac{(w_{0i}(1+r_f)+x_i\ w_{0i}(u-r_f))^{1-\alpha_i}}{1-\alpha_i} f_i + \frac{(w_{0i}(1+r_f)+x_i\ w_{0i}(d-r_f))^{1-\alpha_i}}{1-\alpha_i}(1-f_i) \tag{13.10}$$

The first-order condition w.r.t. x_i, defined as the proportion of investible wealth invested in equity, is therefore:

$$(w_{0i}(1+r_f)+x_i\ w_{0i}(u-r_f))^{-\alpha_i}\ w_{0i}(u-r_f)f_i + (w_{0i}(1+r_f)+x_i\ w_{0i}(d-r_f))^{-\alpha_i} w_{0i}(d-r_f)(1-f_i) = 0$$

with solution independent of w_{0i}, which, with an abuse of notation, can be thought of as investible wealth; this is the specialization of equation (13.9):

$$\left(\frac{1+r_f+x_i(u-r_f)}{1+r_f-x_i(r_f-d)}\right)^{-\alpha_i} = \frac{(r_f-d)(1-f_i)}{(u-r_f)f_i}$$

$$\left(\frac{1+r_f+x_i(u-r_f)}{1+r_f-x_i(r_f-d)}\right)^{\alpha_i} = h_i$$

$$1+r_f+x_i(u-r_f) = h_i^{\frac{1}{\alpha_i}}(1+r_f-x_i(r_f-d))$$

$$x_i(u-r_f)+h_i^{\frac{1}{\alpha_i}}\ x_i(r_f-d) = h_i^{\frac{1}{\alpha_i}}(1+r_f)-(1+r_f)$$

$$x_i\left[u-r_f+h_i^{\frac{1}{\alpha_i}}(r_f-d)\right] = (1+r_f)\left(h_i^{\frac{1}{\alpha_i}}-1\right)$$

Now if $x_i \neq 0$, then

$$x_i = \frac{(1+r_f)\left(h_i^{\frac{1}{\alpha_i}} - 1\right)}{u - r_f + h_i^{\frac{1}{\alpha_i}}(r_f - d)} \tag{13.11}$$

As demonstrated generally in Proposition 1, if agent i is an optimist then $h_i > 1$, $h_i^{\frac{1}{\alpha_i}} > 1$ and hence $x_i > 0$. If i is a pessimist, then $0 < h_i < 1$ and $0 < h_i^{\frac{1}{\alpha_i}} < 1$ and $x_i < 0$. If h_i is a 'martingale', then $h_i = 1$ and the solution is $x_i = 0$.

Now equation (13.11) can be applied in order to solve for equilibrium, thus:

$$nS_0 = \sum I_i x_i = \sum I_i \frac{(1+r_f)\left(h_i^{\frac{1}{\alpha_i}} - 1\right)}{u - r_f + h_i^{\frac{1}{\alpha_i}}(r_f - d)} \tag{13.12}$$

We now see that the equilibrium price is determined by the cross-sectional behaviour of I_i, h_i and α_i. Moreover, I_i depends in turn upon $\overline{n}_i$ and $\overline{m}_i$. However, to compute the equilibrium price taking into account initial wealth dependence adds great extra complexity and little extra clarity, so we ignore this point for the moment.

As an aside, it is noteworthy that equation (13.11) leads to well-defined maximum long and short positions which are independent of α_i and given by the limits of x_i as $h_i \to 0$ and $h_i \to \infty$ as follows:

$$\begin{aligned}\lim_{h_i \to 0} x_i &= \lim_{h_i \to 0} \frac{(1+r_f)\left(h_i^{\frac{1}{\alpha_i}} - 1\right)}{u - r_f + h_i^{\frac{1}{\alpha_i}}(r_f - d)} \\ &= -\frac{(1+r_f)}{u - r_f}\end{aligned}$$

and

$$\begin{aligned}\lim_{h_i \to \infty} x_i &= \lim_{h_i \to \infty} \frac{(1+r_f)\left(h_i^{\frac{1}{\alpha_i}} - 1\right)}{u - r_f + h_i^{\frac{1}{\alpha_i}}(r_f - d)} \\ &= (1+r_f) \lim_{h_i \to \infty} \frac{1}{r_f - d}\left[\frac{(r_f - d)h_i^{\frac{1}{\alpha_i}}}{u - r_f + h_i^{\frac{1}{\alpha_i}}(r_f - d)}\right] \\ &= \frac{(1+r_f)}{r_f - d} \lim_{h_i \to \infty}\left[1 - \frac{u - r_f}{u - r_f + h_i^{\frac{1}{\alpha_i}}(r_f - d)}\right] \\ &= \frac{(1+r_f)}{r_f - d}\end{aligned}$$

Considering typical annual values often used in these calculations, $u = 10\%$, $d = 0\%$, $r_f = 5\%$, we see that these bounds are very large relative to investible wealth, specifically $\pm 2100\%$.

However, it is clear that even if investors take a maximum-sized position (long or short) and the market moves in the opposite direction to their forecast, then their losses are limited to their initial investible wealth. For instance, consider the case of an unanticipated down move when agent i is maximally bullish:

$$\begin{aligned} w_{1i} &= w_{0i}(1+r_f) + x_i w_{0i}(d - r_f) \\ \lim_{h_i \to \infty} w_{1i} &= w_{0i}(1+r_f) + w_{0i}(d - r_f) \lim_{h_i \to \infty} x_i \\ &= w_{0i}(1+r_f) + w_{0i}(d - r_f)\frac{(1+r_f)}{r_f - d} \\ &= 0 \end{aligned}$$

Equation (13.12) as it stands tells us that a redistribution of wealth from optimists to pessimists will lower price and conversely. This raises the question as to when price depends only upon aggregate wealth; that is, under what conditions will the distribution of wealth be irrelevant as long as the aggregate stays the same? Inspection of equation (13.12) allows us to determine conditions that imply the existence of a representative agent which addresses exactly this issue. This is a one-fund separation result as all investors will be holding the same proportion of equity, the representative investor will have aggregate investible wealth and a demand function identical in form to equation (13.12).

Proposition 2: For all agents having differing power utility and beliefs, then if $ln(h_i) = \alpha_i ln(k)$ for all i, where k is a constant greater than 1, we have aggregate demand in the form of equation (13.12) where the aggregate investor, whose investible wealth is aggregate investible wealth, has power utility with risk aversion α and beliefs h which also satisfy the given condition. In particular, our condition requires that all investors be optimists.

Remark 3: The exact aggregation conditions are not fully determined, but if we wished α to equal $\sum \alpha_i$, for example, then h would need to equal $\prod h_i$. If we wished to set α to the average α_i, then h would equal the geometric mean of the h_i. As such, Proposition 2 appears to be a slight generalization of Rubinstein (1974) for the special case of power utility in that we can allow for differing risk aversion *and* beliefs, albeit under restricted conditions.

An interesting point about the aggregation condition is that it requires that agents with high risk aversion should also be highly optimistic. Although nothing in the theory of expected utility seems to help us in terms of assessing the investor's optimism, the result does not sit well with our intuition. An extensive and well-known body of literature

deals with the problem of extracting estimates of investor risk aversion from asset prices. Often, however, the theoretical and empirical frameworks concerned are predicated on the assumption of a single aggregate representative investor and homogeneous beliefs among agents. Remark 3 focuses attention on the perils of drawing strong conclusions about risk aversion from such models, since allowing heterogeneous beliefs means an infinity of representative risk-aversion levels can be observationally equivalent given asset prices.

We note also that price is increasing in $k_i \equiv h_i^{\frac{1}{\alpha_i}}$, so that, *ceteris paribus*, an increase in h_i increases the equilibrium price, an increase in α_i (risk aversion) increases prices if the individual is a pessimist but decreases prices if he is an optimist, as might be expected from Proposition 1.

We can now examine equilibrium situations in which agents differ in wealth, optimism and risk aversion, determine the conditions under which trade occurs, and also who will be buyers and sellers. As a simple illustrative example, consider two agents with an equal initial number of shares and bonds embedded in their equal investible wealth after dealing with their time-0 consumption. They differ only in k_i, and since demand is increasing in k_i the investor with the larger k_i will demand more. We shall assume they cannot both be pessimists so that a sensible equilibrium exists, and also that wealth cannot become negative. We denote the two investors by subscripts A and B.

Net demand for the stock (as a proportion of each investor's identical investible wealth) will be given by

$$x = \frac{(1+r_f)(k_A-1)}{u-r_f+k_A(r_f-d)} + \frac{(1+r_f)(k_B-1)}{u-r_f+k_B(r_f-d)}$$

hence for a strictly positive stock price we require $x > 0$,

that is:

$$\frac{(k_A-1)}{u-r_f+k_A(r_f-d)} > \frac{(1-k_B)}{u-r_f+k_B(r_f-d)}$$

$$(k_A-1)\left[u-r_f+k_B(r_f-d)\right] > (1-k_B)\left[u-r_f+k_A(r_f-d)\right]$$

$$k_A\left[u+d-2r_f+2k_B(r_f-d)\right] > 2(u-r_f)+k_B(r_f-d-u+r_f)$$

$$k_A(1-2\pi)+2k_Ak_B\pi > 2(1-\pi)+k_B(2\pi-1)$$

$$(k_A+k_B)(1-2\pi)+2k_Ak_B\pi > 2(1-\pi)$$

$$(k_A+k_B)\left[\frac{1}{2\pi}-1\right]+k_Ak_B > \frac{1}{\pi}-1$$

$$\left(k_A+\left[\frac{1}{2\pi}-1\right]\right)\left(k_B+\left[\frac{1}{2\pi}-1\right]\right) > \left[\frac{1}{2\pi}-1\right]^2+\frac{1}{\pi}-1$$

$$\left(k_A+\left[\frac{1}{2\pi}-1\right]\right)\left(k_B+\left[\frac{1}{2\pi}-1\right]\right) > \frac{1}{4\pi^2}$$

The lower boundary of the set of (k_A, k_B) pairs which comply with this condition is a hyperbola. It is clear that this will pass through the point *(1,1)* for all values of π. This point represents the risk-neutral beliefs case which we have previously discussed. For further illustration, Figures 13.1 and 13.2 plot examples of the hyperbola for $\pi = 0.9$ and $\pi = 0.1$ respectively. It is interesting to note that when the magnitude of downside move is large relative to the upside (e.g. the $\pi = 0.9$ case), then only a relatively limited degree of pessimism is admissible on the part of one agent if we are to ensure positive prices.

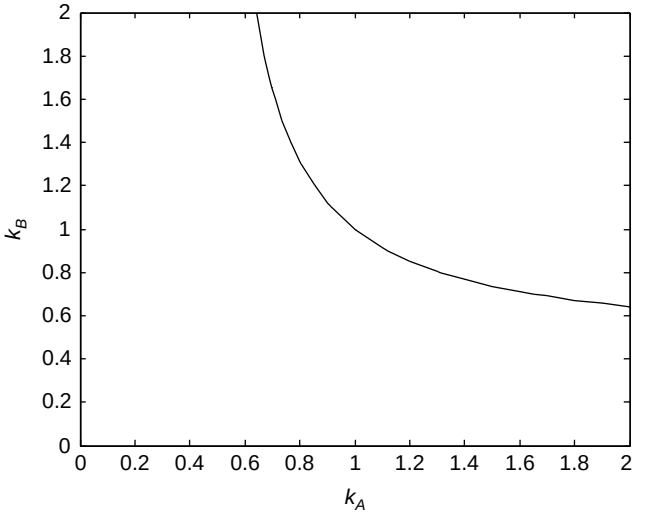


Figure 13.1 Lower boundary of set of (k_A, k_B) pairs resulting in positive price when $\pi = 0.9$; in this case pessimism is relatively hard to accommodate.

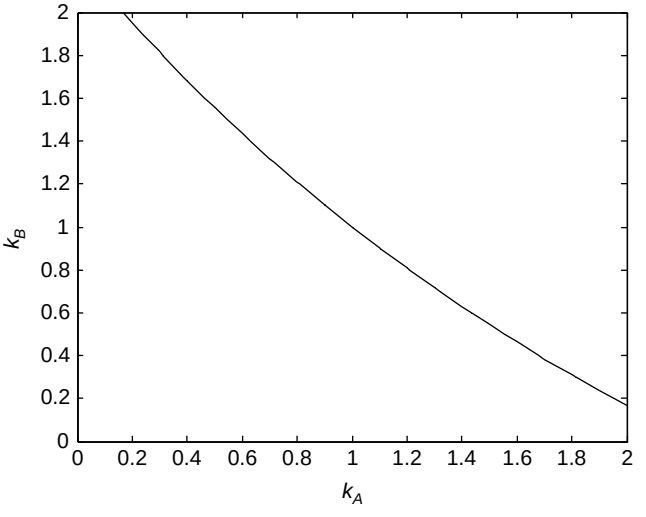


Figure 13.2 Lower boundary of set of (k_A, k_B) pairs resulting in positive price when $\pi = 0.1$; here pessimism is more easily accommodated by an optimist.

13.4 Exponential utility, loss aversion and mixed equilibria

We next consider the case of exponential utility – that is, utility functions with constant absolute risk aversion. In this case, an equilibrium based on different risk aversions is well understood (see, for instance, Rubinstein, 1974), and so there is no gain by considering a binomial economy. However, we could consider the case where some agents have exponential utility whilst others have power utility. The difficulty with this is that their respective wealths are defined over different final values, in that power utility assumes non-negative final wealth whilst exponential utility assumes real final wealth. To advance this further, we need to consider extensions of power utility to include negative final wealth. Instead we shall look at a mixed equilibrium, where one investor is restricted through behavioural considerations whilst the other maximizes power utility. The former, who we shall assume follows the assumptions of prospect theory, will be described next (and referred to as a prospect theorist or prospector, for want of a better term).

The prospector makes investment decisions by distinguishing between utility of *gains* and disutility of *losses*, rather than utility of final wealth indiscriminately. In other words, the prospector cares about *changes* in future wealth relative to some benchmark level rather than its future level *per se*. This can be represented mathematically as a suitably-weighted sum of two expected utility functions, one for gains and the other for losses, where we might, for instance, allow the gain function to display risk aversion while the loss function is risk loving. This formulation allows us to incorporate loss aversion into our economy.

Further, Kahneman and Tversky (1979) stress that the weights associated with utilities need not be objective probabilities. It is this emphasis that allows us to put prospect theory into our framework.

Hence we now consider the problem of a prospect theorist trading with an expected utility maximizer. Both have non-negative wealth; the prospect theorist has a double HARA utility function with a benchmark given by investible wealth invested solely in cash. The expected utility maximizer has a power utility as given by equation (13.10) in Section 13.2. It is immediately apparent without calculation that in a one-period world the investor encumbered by behavioural considerations (i.e. the prospect theorist) suffers no particular disadvantage.

The prospect theorist maximizes the following, after subtracting the relevant benchmark:

$$V_i = \frac{(x_i\ w_{0i}(u - r_f))^{1-\alpha_i}}{1-\alpha_i} f_i - \lambda_i \frac{(-x_i\ w_{0i}(d - r_f))^{1-\beta_i}}{1-\beta_i}(1 - f_i)$$

Here, the parameters $(\alpha_i,\ \beta_i,\ \lambda_i)$ correspond to the coefficients of risk aversion for gains, losses, and a general loss aversion parameter. The first two parameters determine his risk attitude to gains and losses. There are a number of results already known here – for example: if both parameters lie between 0 and 1, the investor is risk averse with respect to gains, and risk loving with respect to losses. Such an outcome is the standard assumption, and was discovered experimentally by Kahneman and Tversky (1979). Typical values are $\alpha_i = \beta_i = .12$ and $\lambda_i = 2.25$. As before, we denote investible wealth by w_{0i}, although strictly it should be I_i. The first-order conditions give us:

$$0 = (x_i\ w_{0i}(u - r_f))^{-\alpha_i} f_i (u - r_f) - \lambda_i (x_i\ w_{0i}(r_f - d))^{-\beta_i}(1 - f_i)(r_f - d) \tag{13.13}$$

Solving equation (13.13), we see that:

$$x_i = \left(\frac{h_i}{\lambda_i}\right)^{\frac{1}{\alpha_i-\beta_i}} \frac{1}{w_{0i}} \frac{(r_f-d)^{\frac{\beta_i}{\alpha_i-\beta_i}}}{(u-r_f)^{\frac{\alpha_i}{\alpha_i-\beta_i}}} \tag{13.14}$$

This solution was derived by Hwang and Satchell (2003). It is worth noting that, despite having constant relative risk aversion with respect to gains and losses separately, our prospect theorist enjoys constant absolute risk aversion overall. Also, since $u > r_f > d$ and $h_i \geq 0$, it is clear that $x_i \geq 0$ as long as λ_i is non-negative – i.e. the prospector will never take an outright short position.

The other investor in the equity market has power utility. His demand for equity is given by equation (13.9), repeated below:

$$x_i = \frac{(1+r_f)\left(h_i^{\frac{1}{\alpha_i}}-1\right)}{u-r_f+h_i^{\frac{1}{\alpha_i}}(r_f-d)}$$

Combining the two, and replacing i by 1 for the power utility investor and 2 for the prospector, we see that the equilibrium price for equity is given by:

$$nS_0 = \sum I_i x_i = w_{01}\frac{(1+r_f)\left(h_1^{\frac{1}{\alpha_1}}-1\right)}{u-r_f+h_1^{\frac{1}{\alpha_1}}(r_f-d)} + \left(\frac{h_2}{\lambda_2}\right)^{\frac{1}{\alpha_2-\beta_2}} \frac{(r_f-d)^{\frac{\beta_2}{\alpha_2-\beta_2}}}{(u-r_f)^{\frac{\alpha_2}{\alpha_2-\beta_2}}} \tag{13.15}$$

It is apparent that, whereas an increase in wealth will scale the demand of investor 1, a change in wealth has no impact on the demand of investor 2. The impact of beliefs on investor 1 will be as described before for power utility. However, an increase in optimism for investor 2 will have a positive or negative impact, depending upon his risk attitudes for gains and losses. For example, if his risk aversion coefficient for losses is less than his risk aversion coefficient for gains, then the equity price is decreasing in loss aversion and increasing in optimism. However, an increase in optimism for the power utility investor always increases the price.

The conclusion we are therefore led to is that an overall increase in external social happiness which, *inter alia*, raises everybody's optimism does not necessarily raise or lower prices. Hwang and Satchell (2003) find that a necessary and sufficient condition for an internal optimum is that $\alpha_2 > \beta_2$, and this condition guarantees that an increase in prospector optimism increases prices whilst an increase in loss aversion decreases prices.

13.5 Conclusions

Despite their apparent simplicity, the various binomial economies described are capable of representing a useful level of richness of agent heterogeneity while remaining analytically tractable. In particular, they present a potentially attractive format for modeling the effects on equilibrium price of differential investor forecasts. Not only can dispersed beliefs be captured, but also significantly different attitudes to risk.

Several interesting directions for future research are apparent, addressing such questions as aggregation conditions, price determination in mixed equilibria, and patterns of trading between agents.

One immediate area of interest is the impact on prices of different assumptions about cross-sectional behaviour. Cross-sectional *volatility* has been measured by practitioners for some time, but cross-sectional *beliefs* have been only roughly measured by consultants' surveys of client optimism. More detailed information of this type would be of particular value in further developing the methods described in this chapter, from both theoretical and applied points of view.

Appendix

We have already noted that equation (13.8) implies that:

$$u_i'(n_iS_0(1+u)+m_i)(u-r_f)f_i = u_i'(n_iS_0(1+d)+m_i)(r_f-d)(1-f_i)$$

By adding $u_i'(n_iS_0(1+d)+m_i)(u-r_f)(1-f_i)$ to both sides of this equation, we obtain:

$$u_i'(n_iS_0(1+u)+m_i)(u-r_f)f_i + u_i'(n_iS_0(1+d)+m_i)(u-r_f)(1-f_i)$$
$$= u_i'(n_iS_0(1+d)+m_i)(u-d)(1-f_i)$$

$$u_i'(n_iS_0(1+u)+m_i)f_i + u_i'(n_iS_0(1+d)+m_i)(1-f_i) = u_i'(n_iS_0(1+d)+m_i)\frac{1-f_i}{1-\pi}$$

$$u_i'(n_iS_0(1+d)+m_i)\frac{1-f_i}{1-\pi} = E_{0i}(u_i'(w_{1i}))$$

$$u_i'(n_iS_0(1+d)+m_i)(1-f_i) = \lambda(1-\pi)$$

A similar method provides:

$$u_i'(n_iS_0(1+u)+m_i)f_i = \lambda\pi$$

where in both cases we define

$$\lambda = E_{0i}(u_i'(w_{1i}))$$

References

Chiarella, C., He, X. and Dieci, R. (2006). Aggregation of heterogeneous beliefs and asset pricing theory: a mean-variance analysis. Research Paper 186, UTS Quantitative Finance Research Centre.

Constantinides, G. M. (1982). Intertemporal asset pricing with heterogeneous consumers and without demand aggregation. *Journal of Business*, 55(2):253–267.

Hara, C., Huang, J. and Kuzmics, C. (2006a). Representative consumer's risk aversion and efficient risk-sharing rules. *KIER Discussion Paper Series*, 620 (May).

Hara, C., Huang, J. and Kuzmics, C. (2006b). Efficient risk-sharing rules with heterogeneous risk attitudes and background risks. *KIER Discussion Paper Series*, 621 (May).

Hwang, S. and Satchell, S. E. (2003). The magnitude of loss aversion parameters in financial markets. Unpublished Discussion Paper, Cass Business School.
Johnstone, D. (2006). Investment as bets in the binomial asset pricing model. University of Sydney Discussion Paper; chapter *Forecasting Expected Returns*.
Kahneman, D. and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, 47(2):263–292.
Rubinstein, M. (1974). An aggregation theorem for securities markets. *Journal of Financial Economics*, 1:225–244.
Varian, H. (1985). Divergence of opinion in complete markets: a note. *Journal of Finance*, 40(1):309–317.

Index